DOCUMENT RESUME

ED 184 561 IR 008 246

AUTHOR Cotton, Ira W.
TITLE Computer Science and Technology: Measurement of Interative Computing: Methodology and Application.
SPONS AGENCY National Bureau of Standards (DOC), Washington, D.C. Inst. for Computer Sciences and Technology.
PUB DATE Jun 79
NOTE 112p.: Ph.D. Dissertation, George Washington University.
AVAILABLE FROM Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402 (Stock No. 003-003-02081-8, $4.00)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Computers; Input Output Devices; *Man Machine Systems; *Online Systems; *Reaction Time; *Research Methodology

ABSTRACT

This dissertation reports the development and application of a new methodology for the measurement and evaluation of interactive computing, applied to either the users of an interactive computing system or to the system itself, including the service computer and any communications network through which the service is delivered. The focus is on the performance of the user and the system in individual interaction sessions with the basic data of interest being the number and rate of characters sent by user and system, and latencies or delays prior to and during transmission by either party. Analysis of the data consists of grouping according to two independent criteria: maximum operating line speed of the terminal and type of application. The data are grouped according to these criteria and cumulative frequency distributions are computed for each of 14 parameters of the model. Non-parametric tests are used to determine the significance of differences in the distributions of different sets of data. The methodology itself is the major contribution of the study, providing a quantitative way to investigate a variety of phenomena associated with interactive computing. The most interesting specific finding from the data collected is the increase in output data length as the terminal speed increases. (Author/RAO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED184561

COMPUTER SCIENCE & TECHNOLOGY

MEASUREMENT OF INTERACTIVE COMPUTING: METHODOLOGY AND APPLICATION

Prepared November 1978

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Ira W. Cotton

Center for Computing Systems Engineering
Institute for Computer Sciences and Technology
National Bureau of Standards
Washington, DC 20234

U.S. DEPARTMENT OF COMMERCE, Juanita M. Kreps, Secretary
Jordan J. Baruch, Assistant Secretary for Science and Technology
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, Director

Issued June 1979

IR008246

FEB 1 3 1980

Reports on Computer Science and Technology

The National Bureau of Standards has a special responsibility within the Federal Government for computer science and technology activities. The programs of the NBS Institute for Computer Sciences and Technology are designed to provide ADP standards, guidelines, and technical advisory services to improve the effectiveness of computer utilization in the Federal sector, and to perform appropriate research and development efforts as foundation for such activities and programs. This publication series will report these NBS efforts to the Federal computer community as well as to interested specialists in the academic and private sectors. Those wishing to receive notices of publications in the series should complete and return the form at the end of this publication.

National Bureau of Standards Special Publication 500-48
Nat. Bur. Stand. (U.S.), Spec. Publ. 500-48, 111 pages (June 1979)
CODEN: XNBSAV

Library of Congress Catalog Card Number: 79-600079

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1979

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402
Stock No. 003-003-02081-8 Price $4.00
(Add 25 percent additional for other than U.S. mailing).

ACKNOWLEDGEMENTS

Dissertations are rarely the product of solely the author's own efforts. Many people help; each deserves thanks. In this case, I am particularly appeciative for the help of my colleague and friend, Dr. Marshall D. Abrams, who originally conceived the measurement tool applied in this research and who was most helpful throughout. Valuable guidance was provided by the members of my committee, including Dr. Ernest H. Forman, Dr. John P. Coyne and especially the Chairman, Dr. Walton E. Smith.* Assistance in data collection and processing was provided by a number of people, including Robert Peterson, Byron Everhart and Bruce Lucas. The illustrations were prepared by Dennis Donovan. Finally, I am deeply grateful to Thomas N. Pyke, Jr., and the National Bureau of Standards for the opportunity to follow this work through to completion.

---

* Note: This report was submitted as a dissertation in partial completion of the requirements for the degree of Doctor of Business Administration at The George Washington University.

TABLE OF CONTENTS

1.0 INTRODUCTION 1

2.0 MEASUREMENT OF INTERACTIVE COMPUTING 5
- 2.1 Measures of Service 5
- 2.2 User Characteristics 7
  - 2.2.1 Input Rate 7
  - 2.2.2 Input Volume and Session Durationn 9
  - 2.2.3 Response Time Requirements 9
  - 2.2.4 Response Time Variability 10
  - 2.2.5 Response Transmission Rate 11
- 2.3 Models of User-Computer Interaction 12
  - 2.3.1 Data Stream Model 12
    - 2.3.1.1 Data Stream Model Parameters 14
    - 2.3.1.2 Model Development 14
  - 2.3.2 Stimulus-Response Model 15
  - 2.3.3 Stimulus-Acknowledgement-Response Model 15
    - 2.3.3.1 Parameters of the SAR Model 17
    - 2.3.3.2 Relationships Among SAR Model Parameters 18
      - 2.3.3.2.1 Compound Delay Times 18
      - 2.3.3.2.2 Transmission Rates 19
- 2.4 Interactive Activity Measurement 20

3.0 DATA COLLECTION 22
- 3.1 Measurement Instrument 22
  - 3.1.1 Network Measurement Machine 23
  - 3.1.2 Data Analysis Package 24
    - 3.1.2.1 The Analysis Model 24
    - 3.1.2.2 Basic Statistics 27
    - 3.1.2.3 Subsets 29
  - 3.1.3 Additional Processing 30
- 3.2 Interactive Environment 31
- 3.3 Measurements Taken 32
  - 3.3.1 Selection of Conversations 32
  - 3.3.2 Subset Assignment 34

4.0 DATA ANALYSIS 35
- 4.1 Review of Data Collected 35
  - 4.1.1 Possible Groupings 35
  - 4.1.2 Data Independence 36
- 4.2 Non-Parametric Tests of Significance 40
  - 4.2.1 Mann-Whitney U-test 40
  - 4.2.2 Kolmogorov-Smirnov Test 41
  - 4.2.3 Ranking and Selection Techniques 42

5.0 EFFECTS OF DIFFERENCES IN TERMINAL SPEED 44
- 5.1 Experimental Design 44
- 5.2 Effects of Terminal Speed on User Performance Parameters 46
  - 5.2.1 Response-Stimulus Delay (Think) Time 46
  - 5.2.2 Stimulus Transmit Time 48
  - 5.2.3 Stimulus Character Count 50
  - 5.2.4 Stimulus Transmission Rate 52

5.3 Effects of Terminal Speed on System Performance Parameters 55
  5.3.1 Stimulus-Acknowledgement Delay Time 55
  5.3.2 Acknowledgement Transmit Time 57
  5.3.3 Acknowledgement Character Count 58
  5.3.4 Acknowledgement Transmission Rate 60
  5.3.5 Acknowledgement-Response Delay Time 61
  5.3.6 Response Transmit Time 63
  5.3.7 Response Character Count 65
  5.3.8 Response Transmission Rate 66
  5.3.9 Stimulus-Response Delay Time 68
5.4 Overall Effects of Varying Terminal Speed 69
  5.4.1 Stimulus Interarrival Time 69
  5.4.2 Line Utilization 71
5.5 Statistical Significance 75

6.0 EFFECTS OF DIFFERENCES IN APPLICATION 78
6.1 Definition of Subsets 78
6.2 Subset Assignment 79
6.3 Tabulation of Summaries 79
6.4 Sample Analyses 79
  6.4.1 Analysis of Particular Parameters 86
  6.4.2 Analysis of Particular Subsets 87

7.0 SUMMARY AND CONCLUSIONS 89
7.1 Review of Research and Findings 89
  7.1.1 Methodology 89
  7.1.2 Results of Analysis by Terminal Speed 90
  7.1.3 Results of Analysis by Application 93
7.2 Applicability of Results 94
  7.2.1 Procurement 94
  7.2.2 User Parameters for System Design 95
  7.2.3 Tuning Interactive Systems 95
7.3 Limitations 96
7.4 Suggestions for Future Work 97

BIBLIOGRAPHY 98

LIST OF TABLES

Table 2-1. User Typing Rates in Jackson & Stubbs Study 8

Table 3-1. Parameter Upper Limits 28
Table 3-2. Summary of Data Recording 32
Table 3-3. Tabulation of "Good" Conversations by Year 33
Table 3-4. Tabulation of "Good" Conversations by Speed 34

Table 4-1. Summary of Message Group Parameters 35
Table 4-2. Data Analysis Alternatives 36

Table 5-1. Number of Message Groups by Parameter and Data Rate 45
Table 5-2. Key Percentiles for Response-Stimulus Delay Time (Seconds) 47
Table 5-3. Key Percentiles for Stimulus Transmit Time (Seconds) 48
Table 5-4. Key Percentiles for Stimulus Character Count 50
Table 5-5. Key Percentiles for Stimulus Transmission Rate (Char/Sec) 52
Table 5-6. Key Percentiles for Stimulus Transmission Rate After Elimination of Paper Tape Input and Single Character Stimuli (Char/Sec) 54
Table 5-7. Key Percentiles for Stimulus-Acknowledgement Delay Time (Seconds) 56
Table 5-8. Key Percentiles for Acknowledgement Transmit Time (Seconds) 57
Table 5-9. Key Percentiles for Acknowledgement Character Count 59
Table 5-10. Key Percentiles for Acknowledgement Transmission Rate (Char/Sec) 60
Table 5-11. Key Percentiles for Acknowledgement-Response Delay Time (Seconds) 62
Table 5-12. Key Percentiles for Response Transmit Time (Seconds) 63
Table 5-13. Key Percentiles for Response Character Count 66
Table 5-14. Key Percentiles for Response Transmission Rate (Char/Sec) 67
Table 5-15. Key Percentiles for Stimulus-Response Delay Time (Seconds) 69
Table 5-16. Key Percentiles for Stimulus Inter-arrival Time (Seconds) 70
Table 5-17. Mean Values of Components of Stimulus Inter-Arrival Time 71
Table 5-18. Line Utilization of 110 BPS Conversations 72
Table 5-19. Line Utilization of 150 BPS Conversations 73
Table 5-20. Line Utilization of 300 BPS Conversations 74
Table 5-21. Results of Non-Parametric Tests Between 110 BPS and 300 BPS for All Parameters Using All Data (Number of Observations Given as Given in Table 5-1) 76
Table 5-22. Results of Non-Parametric Tests Between 110

7

BPS and 300 BPS for All Parameters Using Medians of Each Message Group (Number of Observations as Given in Table 2-3) 77

Table 6-1. Subset Assignment Summary for 110 BPS Conversations 80
Table 6-2. Subset Assignment Summary for 150 BPS Conversations 80
Table 6-3. Subset Assignment Summary for 300 BPS Conversations 81
Table 6-4. Key Percentiles for All Parameters for 300 BPS Conversations, Grouped by Application 82
Table 6-5. Stimulus Transmission Rate and Stimulus Character Count by Application, Sorted by Increasing Values of Stimulus Transmission Rate 87

LIST OF FIGURES

Figure 1-1. Dissertation Outline 3

Figure 2-1. Data Stream Model 13
Figure 2-2. SAR Model 17

Figure 3-1. Sample Conversation Record 26

Figure 4-1-a. Distribution of Autocorrelation Coefficients for Response-Stimulus Delay Time for All Conversations 38
Figure 4-1-b. Distribution of Autocorrelation Coefficients for Stimulus Transmit Time for All Conversations 38
Figure 4-1-c. Distribution of Autocorrelation Coefficients for Stimulus Character Count for All Conversations 38
Figure 4-1-d. Distribution of Autocorrelation Coefficients for Stimulus Transmission Rate for All Conversations 38
Figure 4-1-e. Distribution of Autocorrelation Coefficients for Stimulus-Achnowledgement Delay Time for All Conversations 38
Figure 4-1-f. Distribution of Autocorrelation Coefficients for Acknowledgement Transmit Time for All Conversations 38
Figure 4-1-g. Distribution of Autocorrelation Coefficients for Acknowledgement Character Count for All Conversations 38
Figure 4-1-h. Distribution of Autocorrelation Coefficients for Acknowledgement Transmission Rate for All Conversations 38
Figure 4-1-i. Distribution of Autocorrelation Coefficients for Acknowledgement-Response Delay Time for All Conversations 39
Figure 4-1-j. Distribution of Autocorrelation Coefficients for Response Transmit Time for All Conversations 39
Figure 4-1-k. Distribution of Autocorrelation Coefficients for Response Character Count for All Conversations 39
Figure 4-1-l. Distribution of Autocorrelation Coefficients for Response Transmission Rate for All Conversations 39
Figure 4-1-m. Distribution of Autocorrelation Coefficients for Stimulus-Response Delay Time for All Conversations 39
Figure 4-1-n. Distribution of Autocorrelation Coefficients for Stimulus Inter-Arrival Time for All Conversations 39

Figure 5-1. Distribution of Response-Stimulus Delay Time for All Message Groups, According to Terminal Speed 47

Figure 5-2. Distribution of Stimulus Transmit Time for All Message Groups, According to Terminal Speed 49
Figure 5-3. Distribution of Stimulus Character Count for All Message Groups, According to Terminal Speed 51
Figure 5-4. Distribution of Stimulus Transmission Rate for All Message Groups, According to Terminal Speed 53
Figure 5-5. Distribution of Stimulus-Acknowledgement Delay Time for All Message Groups, According to Terminal Speed 56
Figure 5-6. Distribution of Acknowledgement Transmit Time for All Message Groups, According to Terminal Speed 58
Figure 5-7. Distribution of Acknowledgement Character Count for All Message Groups, According to Terminal Speed 59
Figure 5-8. Distribution of Acknowledgement Transmission Rate for All Message Groups, According to Terminal Speed 61
Figure 5-9. Distribution of Acknowledgement-Response Delay Time for All Message Groups, According to Terminal Speed 62
Figure 5-10. Distribution of Response Transmit Time for All Message Groups, According to Terminal Speed 64
Figure 5-11. Distribution of Response Character Count for All Message Groups, According to Terminal Speed 65
Figure 5-12. Distribution of Response Transmission Rate for All Message Groups, According to Terminal Speed 67
Figure 5-13. Distribution of Stimulus-Response Delay Time for All Message Groups, According to Terminal Speed 68
Figure 5-14. Distribution of Stimulus Inter-arrival Time for All Message Groups, According to Terminal Speed 70

Figure 7-1. User Satisfaction as a Function of Terminal Speed 92

MEASUREMENT OF INTERACTIVE COMPUTING: METHODOLOGY AND APPLICATION

Ira W. Cotton

This dissertation addresses the measurement of interactive computing, including both the computer system providing service and the users demanding and receiving it. The focus is on the performance of the user and the system in individual interaction sessions (rather than on the performance of the system under varying conditions of load). A new measurement tool developed at the National Bureau of Standards is employed to record a large number of individual interactive sessions over a period of three years. The basic data of interest are the number and rate of characters sent by user and system, and latencies or delays prior to and during transmission by either party. These data are fit to a model of user-computer interaction which distinguishes between stimuli from the user, acknowledgements from the system (which only indicate that a service request has been received) and responses from the system (which contain meaningful information).

Analysis of the data consists of grouping according to two independent criteria: 1) maximum operating line speed of the terminal (either 10, 15 or 30 characters per second); and 2) type of application (for each individual service request). The data are grouped according to these criteria and cumulative frequency distributions are computed for each of 14 parameters of the model. Non-parametric tests are used to determine the significance of differences in the distributions for different sets of data.

The methodology itself is the major contribution of the study, providing, as it does, a quantitative way to investigate a variety of phenomena associated with interactive computing. The most interesting specific finding from the the data collected is the increase in output data length as the terminal speed increases.

Key Words: Computer performance evaluation; human factors; interactive computing; man-machine interaction; performance measurement; timesharing

1.0 INTRODUCTION

This dissertation addresses the measurement of interactive computing -- measurement both of the computer system providing the service (including the communications system delivering it) and of the humans using it. Interactive computer systems, by their very nature, include the users as a system component whose performance affects the overall system. Traditional computer system measurement has failed to consider the needs of individual computer system users (focusing instead on overall performance measures), nor has it considered the impact of user performance on computer system performance.

In this dissertation, a measurement approach is described and applied that does address both these measurement needs for interactive computing. A new measurement tool developed at the National Bureau of Standards is employed to measure actual interactive sessions or conversations on a timesharing computer system over a period of three years. Only successful and representative conversations are considered in this study; aborted or otherwise erroneous conversations were discarded. The basic data of interest for the "good" conversations are the number and rate of characters sent by user and system, and latencies or delays prior to and during transmission by either party. These data are fit to a model of user-computer interaction which distinguishes between stimuli or input from the user, acknowledgements from the system (which only indicate that a service request has been received) and responses from the system (which contain meaningful information).

Statistical procedures are developed and employed to analyze the significance of the results when the data are fit to this model of user-computer interaction and then grouped according to various criteria. Analysis of the data consists of grouping according to two independent criteria: 1) maximum operating line speed of the terminal (either 10, 15 or 30 characters per second); and 2) type of application, as indicated by the software subsystem invoked by each individual service request in the conversation. The data are grouped according to these criteria and cumulative frequency distributions are computed for each of 14 parameters of the model, both for all observations in the group, and for the medians of the set of all applicable observations in each conversation (as a way of batching the data to eliminate any evident serial correlation between successive observations in a conversation). Non-parametric tests are used to determine the significance of differences in the frequency distributions for different sets of data.

In the case of grouping by terminal speed, significant information is obtained characterizing both user and system performance. In the case of grouping by application, the data are not analyzed so exhaustively; rather, the potential types of analyses are illustrated and motivated. Other applications for this measurement approach are also suggested.

The methodology itself if the major contribution of the study, providing, as it does, a quantitative way to investigate a variety of phenomena associated with interactive computing. The most interesting specific finding from the data collected is the increase in output data length as the terminal speed increases. This observed increase, over the range of speeds studied, contrasts with no observed increase in a previously reported study with a range of speeds higher than those in this study. These results lead to the postulation of an upper limit on the utility of terminal speed to users (with current modes of interaction).

Figure 1-1 serves to outline the organization of the thesis in graphical form. Chapter 1 is the current introduction and overview.

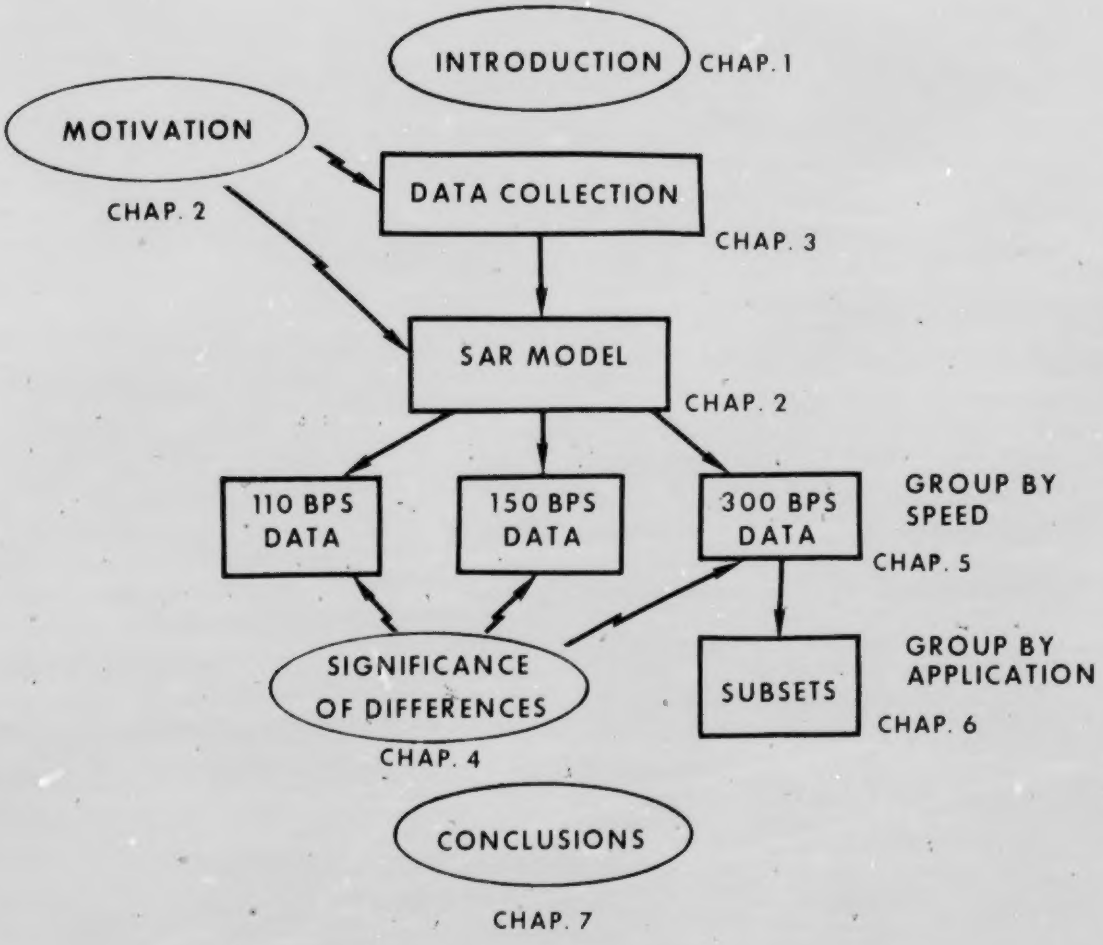

INTRODUCTION
CHAP. 1
MOTIVATION
CHAP. 2
DATA COLLECTION
CHAP. 3
SAR MODEL
CHAP. 2
110 BPS DATA
150 BPS DATA
300 BPS DATA
GROUP BY SPEED
CHAP. 5
SIGNIFICANCE OF DIFFERENCES
CHAP. 4
SUBSETS
GROUP BY APPLICATION
CHAP. 6
CONCLUSIONS
CHAP. 7

Figure 1-1. Dissertation Outline

Chapter 2 provides background on the measurement of interative computing. The difference between traditional measurement techniques that focus on overall system performance and new techniques that focus on the service provided to and the behavior of individual users is clarified. The importance of the user in interactive systems is stressed, followed by a review of the relevant system characteristics of users considered as system components. Several models of user-computer interaction are presented, including a recently developed one that is used in the analysis of the data collected in the study. The chapter concludes with a consideration of the data collection requirements of the model.

Chapter 3 deals with the actual data collection procedures for this investigation. A novel data collection instrument, the National Bureau of Standards' Network Measurement Machine, is described along with the data analysis software available for use with it. This instrument was used to measure interactive use of the NBS Univac 1108 over a three year period. Only "normal" recorded conversations are analyzed. The volume of data collected and the screening criteria are described.

Chapter 4 deals with the analysis procedures for the data after it has been collected. Two different ways of grouping the data are described: (1) according to the speed of the interactive terminal used (110 bits per second, 150 bps or 300 bps), and (2) according to the software subsystem invoked by the user at each stage in the interactive session. The issue of data independence between succeeding observations in the same conversation (serial or auto-correlation) is addressed and test results are presented. The skewed nature of the data makes it impossible to apply statistical tests based on the normal distribution. Two non-parametric tests, the Mann-Whitney U-test and the Kolmogorov-Smirnov test, appear to be suitable to analyze data of this type. The bases of these tests are described.

In Chapter 5 the results of analyzing the data according to the selected model of user-computer interaction, grouped according to terminal speed, are presented. Median and 90-percentile statistics are tabulated for each of fourteen parameters for the different speed terminals, and the results are discussed. The differences between the parameter distributions for the 110 bps and 300 bps data are tested for significance using the Mann-Whitney U-test and the Kolmogorov-Smirnov test and the results presented.

In Chapter 6 the results of analyzing the data according to the same model, but grouped according to software system invoked by the user, are presented. Median and 90-percentile statistics are tabulated for each of the fourteen parameters for the different software systems, and the results are discussed.

Chapter 7 concludes the study with a summary of methodological contributions and empirical findings, discussion of areas of application, identification of limitations and suggestions for future work in the area of interactive computing measurement. Primary areas of possible application include benchmarking in system procurements, gathering user parameters for use in specific system design, and tuning interactive systems. In fact, the approach has already been applied in procurement situations.

2.0 MEASUREMENT OF INTERACTIVE COMPUTING

In recent years, increasing numbers of people have begun to use computers through interactive terminals in a conversational mode. Rather than submitting jobs in a batch mode and waiting hours (or days) for the results, users interact with the computer continuously on a transaction-oriented basis. This trend gives every evidence of continuing, and it appears that this will be the predominant form of computer access in the future.

The history of interactive systems has been marked by an inability to come to grips with the design problems raised by the inter-relationships of the many diverse components that constitute such systems. As interactive systems have become larger and more ambitious, so too have the failures become larger and more notable (e.g., reservation systems that simply can not handle the load). Thus, there is a need for quantitative data describing the inter-relationships of the various components of interactive computer systems and relating the performance of these components to the performance of the overall system. This data is needed by system designers and implementers for use in setting design goals, as well as by users and system procurers for use in specifying requirements.

In this chapter, we consider the differing goals of traditional system measurement and evaluation as compared to the measurement and evaluation of interactive systems. We see that the focus is, of necessity, quite different. System managers seek to optimize the overall operation of the system for maximum efficiency, while individual users are primarily concerned with the level of service they each receive. Frequently, tradeoffs must be made between efficiency and level of service to individual users. Also, the performance of the users themselves impacts the functioning of the computer system. Thus, there is a need to consider the user as a system component for which "operating data" is required. Hence, we review the relevant characteristics of human operators. We then describe several models that can be employed to characterize the interactive situation, including one that will be used for the analysis of the data collected in this study. We conclude with an introduction to the measurement techniques suitable for investigating interactive computer usage.

2.1 Measures Of Service

Much attention has been directed in recent years at measurement and related quantitative evaluation techniques as an important part of the selection and improvement of the hardware and software of computer systems. Both hardware and software "tuning" of the system can be undertaken in response to the information gathered. The overall results of such tuning may be quite dramatic increases in system throughput.

The basis for measuring success in performance evaluation of this type is some measure of the cost-effectiveness of the system before and after the changes. The cost-effectiveness calculations usually are based on the time to run some given set of jobs (a "benchmark") and the cost (purchase or monthly rental) of the particular configuration. Thus, cost-effectiveness is really a ratio of throughput to cost, and the measurement analyst is presumed to be indifferent to the choice between a reduction in cost for identical throughput, or increased throughput for identical cost. The time to process the benchmark is important only as a measure of throughput.

Individual users, however, may not be indifferent to this tradeoff of cost versus time. For them, service is the key factor, not throughput. In contrast to the evaluation of efficiency, which is concerned with the time and cost to run a group of jobs, service evaluation is concerned with the time and cost to run each individual job. Actually, the times of concern are of different types. Efficiency is concerned only with running or execution times, while service evaluation is concerned with total elapsed time. (In most multi-programming systems, the elapsed time for a job is considerably greater than the run time). Thus, the evaluation of efficiency is based on the measurement of the internal functioning of a computer system as a whole, rather than of its external manifestations, taken individually.

The goals of improved efficiency and improved service may well be at odds with one another. An improvement in internal performance does not necessarily imply an improvement in service. Indeed, the opposite may be true. Frequently, it is possible to improve service only at the expense of efficiency.

For measuring service, the methodology of cost-benefit analysis is more appropriate than that of cost-effectiveness (Cotton, 1974; Streeter, 1972). Cost-benefit analysis recognizes that jobs may have a time-value, and that a simple average of throughput is an inadequate measure. In a sense, measures of service also must be concerned with the standard deviation of run or response time. When measuring performance, jobs run more quickly than average cancel out the effect of more slowly run jobs; in measuring service there may be only two types of jobs - acceptable and unacceptable. No further merit is ascribed to jobs run more rapidly than what is deemed "acceptable"; there is no averaging between jobs run exceptionally well and exceptionally poorly.

The most common performance measure applied to conversational systems is "response time." This measure is most frequently defined to be the elapsed time from the end of the user's input to the beginning of the system's response (National Bureau of Standards, 1978).*

---

* Other measures of the system's responsiveness have been proposed (Abrams, Lindamood and Pyke, 1973).

The general goal of most system designers has been to achieve the best (shortest) response time possible. Too often, however, systems have been designed without a particular response time goal in mind, possibly because it was not known what response time was needed. This determination is most properly the domain of engineering-oriented psychologists, some of whom have begun to address the problem in recent years. In the following section, we review the functional characteristics of users that are relevant to the analysis of interactive systems.

2.2 User Characteristics

Considering the importance of the user in the design and performance of interactive computing systems, surprisingly little is known about user characteristics. The traditional approach of the human factors engineer is to regard the user as another system component and to seek to optimize the total system in considerations of the capabilities and limitations of the human component (Van Cott, 1972). Too often, though, interactive systems have been assembled by hardware and software specialists without any assistance from such human factors engineers.

The role of the user in interactive systems is both as a source and sink for information. As an information source, the user is characterized by the rate at which information is entered (including both mean and burst characteristics), by the latency of new inputs following receipt of a computer response, and by the nature of the information entered (e.g., the software system used). The role of the user as an information sink is only measured indirectly in terms of successive latencies and changes in source information content. User characteristics as a sink are such as to place demands on the system in terms of its responsiveness and the volume of response information.

2.2.1 Input Rate -

User typing capabilities have been studied extensively as part of the human factors of data entry devices and procedures (Seibel, 1972). For high-volume entry of redundant data such as English text on typewriter-like keyboards, speed test rates of 60 words (300 characters) per minute are quite common, with an upper limit of about 100 w.p.m. in production situations and "championship" speeds approaching 150 w.p.m. (Devoe, 1967). This range of 60 to 150 w.p.m. corresponds to a range of 5 to 12.5 keystrokes per second. (These are sustained rates which can be exceeded by "burst" rates for brief intervals). Unskilled typists typically enter text at about 1 stroke per second.

For keypunching, a rate of 170 characters per minute (2.8 characters per second) is reported as a good estimate of the mean rate of daily entry for time spent at the machine (based on data from many different data entry jobs at several different keypunch installations) (Klemmer, 1960). No consistent differences are reported between alphanumeric and straight numeric keypunching.

For input to interactive computer systems, the data most widely circulated is based on users of terminals constrained to a relatively low maximum rate of 10 or 15 characters per second. While the models of terminals used in the sample were not explicitly identified, it may be inferred from the date of the study (late 1960's) and the speed of the terminals that they were tele- typewriters, most likely Model 33 or 37 Teletypes. Such terminals have a keyboard layout similar to that of office typewriters but quite different in key type and "feel" from conventional typewriters.

From the data on high-volume typing given above, it can be seen that experienced typists can often exceed the capacity of the terminal to accept keystrokes, especially for short bursts of data entry. This data should not be applied blindly, since, as just noted, the keyboard "feel" of a teletypewriter is not the same as a conventional typewriter and does, in fact, reduce the achievable maximum typing rate of the operator. In addition, the loading of the computer system, and consequently its failure to acknowledge input within the required time (see section 2.1.3) or the limited size of its buffers for yet unprocessed input characters, may also serve to constrain the maximum possible rate of user input.

Bryan (1967) reported a median input rate of about 0.6 characters per second and a mean of about 1.2 characters per second for timesharing users on interactive "typewriters" with a maximum input/output rate of 15 cps.

The following results may be derived from data reported by Jackson and Stubbs (1969) of Bell Telephone Laboratories for three different types of interactive systems (only mean rates were given):

Terminal Speed (chars./sec.)	10	10	15
Application	Scientific	Scientific	Business
Load	Moderate	Heavy	Moderate
User Input Rate (chars./sec.)	.48	.27	1.22

Table 2-1. User Typing Rates in Jackson and Stubbs Study

These rates include the effect of user delay after computer response (user think time) as well as inter-character delays after initiation of user input.

2.2.2 Input Volume And Session Duration -

The volume of input to be expected from each user is important to system designers since it, along with the rate, is another factor in the demand placed on the computer/communications system. Session duration, or holding time, is of particular concern to designers of communications systems where facilities are dedicated to a user for the duration of a session (or call) and where billing may be on a per call rather than a metered basis.

Bryan (1967) reported a median input length of about 8 characters with a mean of about 13 characters. The distribution of input lengths for this study appears to be a negative exponential when plotted as percentages. (In contrast, median and mean output line lengths were about 22 and 32 characters, respectively, and were more uniformly distributed.) The average user entered 82 such lines in a session lasting 46 minutes (median session duration was about 20 minutes).

Jackson and Stubbs (1969) reported the mean number of input characters per interaction for each of the three systems listed in Table 2-1 above as 9.2, 10.7 and 13.8, respectively. (Response length, only cited as a mean for the three systems, was 47 characters.) Average holding times for the three systems were 17, 34 and 21 minutes, respectively.

2.2.3 Response Time Requirements -

The seminal work in the area of response time requirements is that of R. Miller (1968). The author, a behavioral scientist, attempted to list and define the different classes of operator activity at interactive terminals along with the response time requirement for each type of activity. Sixteen specific categories of activities are discussed, which can be summarized by three general classes of activities (Cotton, 1969).

The first activity class is the input of data to the system or control activation of some function through a keyboard or other entry device. An immediate response of no longer than 0.1 - 0.2 seconds is said to be required for this class in order to indicate acceptance by the system. The second class is characterized by a user engaged in high-intensity "brainstorming" requiring the ready access of data from the user's own short-term memory. Applications might incude short searches of files or manipulation (editing) of data. Such activity is said to require no longer than about a 2 second response in order that the chain of thought not be broken. The final class includes those activities which complete a subjective (sub)task or (sub)purpose, called a "closure." Users are said to tolerate more extended delays (up to about 15 seconds) following such an activity completion or closure, than in the process of achieving a closure. No activities are cited for which the user will tolerate response delays longer than about 15 seconds.

These suggestions have been widely used in developing performance specifications for interactive systems and, in fact, may even have influenced system design itself, insofar as responses to different types of activities can be provided by different system components (such as local processors in "intelligent terminals"). However, it is unfortunate that the discussion, though convincing, was not accompanied by any supporting empirical evidence.

There is some data showing the effects of increasing system response time -- e.g., on subsequently increased user response times (Boies, 1974), and on (predicted) decreasing user acceptability ratings (Carbonnel, Elkind, & Nickerson, 1968). Lancaster and Fayen (1973) review similar considerations of system response time from the point of view of information retrieval.

2.2.4 Response Time Variability -

The degree of attention that system designers traditionally have devoted to response time alone may be questionable. L. A. Miller and Thomas (1976) suggest that in view of "the profound technological improvements (hardware and software) that ... permit greatly improved system response times, concerns about the absolute magnitude of delays may no longer be warranted." As Carbonell et. al. (1968) and others have observed, it is the variability of delays, not their magnitude, which is often most distressing to the user. Unfortunately, the response provided by computer systems is quite often subject to considerable variability both within and across sessions, primarily, but not exclusively, due to the varying load placed on the system (Bell, 1974).

Investigation into the performance of console operators as a function of response time variability has a basis in the concern of experimental psychologists with the reaction time of subjects in various stimulus-response settings. L. H. Miller (1976) presents a brief review of the literature in this area, which generally show that the reaction time increases as the variability increases (Mackworth, 1970; Mostofsky, 1970; Davies, 1969). Miller deduces a theory of reaction time which involves the expectancy of the next signal on the part of the subject: "It appears that subjects perform best when the next signal occurs at a time approximately equal to the mean inter-arrival rate of all previous signals. As the arrival of the next signal occurs significantly before or significantly later than this mean arrival rate of previous signals, the subject's arousal is decreased and response suffers."

Boehm, Seven and Watson (1971) studied the effects of response time variability by introducing additional delay to one group of users in order to reduce the variability. This group received a longer mean delay, but smaller variability in response time than a control group working on an identical set of test problems. The long delay, low variability group produced better or faster solutions to the problems than the control group, though they expressed less satisfaction with the computer system. The authors suggested that the extra delay gave the subjects time to reflect on what they were doing and to formulate their solution strategies more carefully.

2.2.5 Response Transmission Rate -

The response rate of a computer, measured in terms of characters per second, has been found by all the studies previously cited (e.g., Bryan, 1967; Jackson and Stubbs, 1969) to be significantly faster than the user input rate (perhaps even by an order of magnitude). The burst rate in most cases approaches the limit of the terminal/communications line capacity. These data are as important as user input data rates in the design of a communications system which, after all, must carry traffic in both directions; however, in this section we are less concerned with the design implications of the response transmission rate than with its impact on the user.

It seems to be generally accepted as an article of faith that users prefer more rapid response transmissions from the computer. Empirical evidence to support this has been lacking, though.

There also has been a realization that there is an upper bound on the improvement in user performance and/or satisfaction resulting from increased response transmission rates. Jackson and Stubbs (1969) conjectured that at higher transmission rates, both the user think time and the volume of information requested would increase. The premise for the first increase (think time) is based on the observation that "in at least some instances, the user utilizes computer send time to read the output he receives. Hence, if the computer outputs the same number of characters in a much shorter time interval, the user may increase his think time in order to do the same amount of reading and thinking." As a result, Jackson and Stubbs suggest the existence of an upper bound on computer transmission rate "beyond which decreases in computer send time are matched by equal increases in user think time." As regards output volume, the authors only suggest that it will "naturally increase."

L. H. Miller (1976) recently provided evidence to support the first of Jackson and Stubbs' conjectures. In this study, the effects of varying CRT display rates and output delays upon user performance and attitudes in a series of message retrieval tasks were evaluated experimentally. The results supported the "somewhat surprising" (to someone evidently unfamiliar with the Jackson and Stubbs paper) conclusion that doubling the display rate from 120 characters per second to 240 cps produced no significant performance or attitude changes. The study did not include experiments at the lower rates of 10, 15 or 30 cps.

2.3 Models Of User-Computer Interaction

Complex phenomena are often best understood in terms of models. Model making and testing is a fundamental part of the "scientific method" which we are seeking to apply to our investigation of user-computer interaction. For the purposes of this discussion it is sufficient to note two points about conceptual models:

(1) The simplest model that incorporates all the observations describing a phenomenon is preferred.

(2) When interpretation of the data according to a model produces results that are at variance with reality, it is time to change the model.

In this section, we examine the model on which most of the current data used by data communications systems analysts is based. We then describe a simpler model that was proposed for a follow-on set of investigations intended to update the data, and show how that model had to be modified when interpretation of the data led to anomalous results. The modified version is presented in detail, since it was used to interpret all the new empirical data collected in this investigation.

2.3.1 Data Stream Model -

The first published analytic model of the communications process between a multiaccess computer and a user at a remote terminal was developed at the Bell Telephone Laboratories (Jackson and Stubbs, 1969; Fuchs and Jackson, 1970; Dudick, Fuchs and Jackson, 1971). The model describes the communications process in terms of random parameters based on times between characters transmitted through the communications network. All of the parameters are measurable at either the communications interface to the computer or to the terminal.

The model was developed by a communications carrier organization to focus on the user-computer communications process and to show how the characteristics of the computer and of the user affect communications requirements. (It is also used to study the converse, i.e., how the constraints of the communications medium affect the user and the computer). As such, the model is formulated in terms of detailed information on the timing relationships within a call or interactive session which is used to develop an expression for the holding time or duration for the call.

Figure 2-1 illustrates the "data stream model", as it is called. A "call" is the total time period during which the user sends and receives characters, including any idle time between either user or computer characters. The model as developed is a half-duplex model, that is, it does not account for the simultaneous sending and receiving of characters.* A period of transmission by either the user or the computer is called a burst segment. By definition, burst segments begin at the end of the last character sent by the other party.

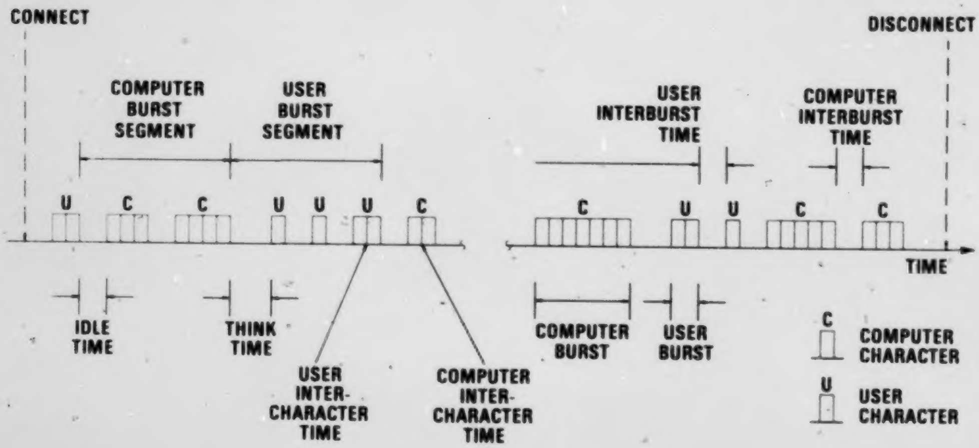

CONNECT
DISCONNECT
COMPUTER BURST SEGMENT
USER BURST SEGMENT
USER INTERBURST TIME
COMPUTER INTERBURST TIME
U
C
C
U
U
U
C
C
U
U
C
C
TIME
IDLE TIME
THINK TIME
USER INTER-CHARACTER TIME
COMPUTER INTER-CHARACTER TIME
COMPUTER BURST
USER BURST
C
COMPUTER CHARACTER
U
USER CHARACTER

Figure 2-1. Data Stream Model

Within a given burst segment, there are periods of line activity and of line inactivity (idle time). The idle time before the first user character in each burst segment is identified as the "think" time. The remaining inactive periods within a burst segment are called intercharacter and interburst times. A burst is defined as a string of consecutive characters each separated by less than one-half character width. Obviously, the number of bursts in a burst segment is less than or equal to the number of characters in that segment.

---

* Jackson and Stubbs assert that "simple modifications" to the model would accommodate full duplex operation. The problem may be more complex than they realized, however. See, for example, the discussion by Abrams and Cotton (1975, pp. 13-14).

2.3.1.1 Data Stream Model Parameters -

The twelve parameters of the data stream model (which are all used in the expression for the total holding time of a call) are as follows:

S = number of burst segments per call
T = think time (user)
I = idle time (computer)
B^u = user interburst time
B^c = computer interburst time
N^u = number of bursts per user burst segment
N^c = number of bursts per computer burst segment
M^u = number of characters per user burst segment
M^c = number of characters per computer burst segment
C^u = user intercharacter time
C^c = computer intercharacter time
W = character width (in seconds)* $[W^u = W^c]$

2.3.1.2 Model Development -

An expression for the holding time of a call was developed by summing expressions for eight individual call components: (1) total user character time; (2) total computer character time; (3) total user intercharacter time; (4) total computer intercharacter time; (5) total user interburst time; (6) total computer interburst time; (7) total think time; and (8) total idle time. The resulting expression is a complex formula with the eight terms each summed over the appropriate number of burst segments (approximated by S/2 each for user and computer). Knowing the distributions for each of the twelve parameters in the model, the authors state that it is theoretically possible, but computationally prohibibitive, to solve directly for the distribution of holding time. Instead, they solve for the mean value of the holding time which they obtain, under the assumptions that the random values are stationary and mutually independent, by taking the expected value of the expression. This result is expressed as the sum of the following four component parts, each having its own functional significance (the symbol for each variable represents its mean value):

a. user send time (the total amount of time during which user characters are being transmitted)

$$= (S/2)(N^uM^uW^u) \qquad (1)$$

b. computer send time (the total amount of time during which computer characters are being transmitted)

---

* Character width = 1/(character transmission rate)

$$= (S/2)(N^cM^cW^c) \qquad (2)$$

c. user delay (the sum of all inactive periods during user burst segments)

$$= (S/2)[T + B^u(N^u-1) + N^uC^u(M^u-1)] \qquad (3)$$

d. computer delay (the sum of all inactive periods during computer burst segments)

$$= (S/2)[I + B^c(N^c-1) + N^cC^c(M^c-1)] \qquad (4)$$

2.3.2 Stimulus-Response (SR) Model -

For many applications, the data stream model is unnecessarily complex. For example, the identification of individual bursts within burst segments from either the user or the computer is unnecessary if the conversation is to be viewed simply as a sequence of individual transations whose start and endpoints only need be identified. This approach was taken by researchers at the National Bureau of Standards in the two models developed there (Abrams and Cotton, 1975).

The first of these, the stimulus-response (SR) model, is the simplest model that can be defined of user-computer interaction. In this model the human is considered to be in control. User inputs to the computer are termed "stimuli"; outputs to the human are termed "responses." The simplest pattern of computer usage is a transaction in which the user issues a stimulus and waits until the computer returns a response. It may be assumed that the next stimulus is dependent upon the contents of the preceding response. This transaction, or stimulus-response pair, is known as a "message group."

2.3.3 Stimulus-Acknowledgement-Response (SAR) Model -

Interpretation of data from certain systems according to the stimulus-response model led to the anomaly that computer conversations which left the human user with a subjective evaluation of very slow response appeared to have very fast response when analyzed (Abrams, Lindamood and Pyke, 1973). Study of the data revealed that communications conventions had not been considered adequately. The American Standard Code for Information Interchange (ASCII) specifies the format effector interpretation of the carriage return (CR), but not its conventional use in conversational applications. Many computers employ the convention that a line constitutes a unit of input and that a line is terminated by a CR. Since the CR (only) returns the print position to the first column of the current line, the computer issues a line feed

(LF) following receipt of a CR to move the print position to the following line. (Some computers require that the sequence CR LF terminate the line, thus obviating the problem). Other control characters may accompany the LF, but their presence generally is not even recognized by the user since they cause no printing or motion of the print mechanism.

According to the stimulus-response model, these non-printing characters constitute the beginning of the response; by subjective human judgment they do not. Their presence is important, however, in that they provide feedback which reassures the user that the computer is still functioning. Whenever the user types a CR, he is reassured that the computer is still serving him by the action of the LF. Nevertheless, this LF does not constitute a meaningful response to the stimulus. In psychological terms, the LF provides partial closure (R. Miller, 1968), but the user is still waiting for complete closure to be provided by the response.

To account for this anomaly, the stimulus-response model was modified. A new state was introduced called the "acknowledgement"; this model is accordingly called the "stimulus-acknowledgement-response" (SAR) model. Operationally, the acknowledgement was first defined as all the initial non-printing characters which are output by the computer following a stimulus. As in the case of the LF, the acknowledgement provides a partial closure by reassuring the user of the system's continuing ability to serve him.

Since the concept of an acknowledgement is based on psychological considerations, its content must be defined subjectively. Once defined, the acknowledgement can be recognized by an algorithm. There are two extensions to the definition of the ackowledgement. First, the acknowledgement is viewed as being time-dependent and may therefore be terminated by the lapse of some specified time period during which there is no transmission from the computer. Second, the response may be defined as beginning with the first meaningful character printed by the computer. In particular, a fixed heading on all output may be considered as part of the acknowledgement rather than as part of the response.

Some interactive computer systems appear to construct an entire output buffer before initiating any output to the user. The buffer contents are then all transmitted to the user at the maximum network transmission rate. In such a case, even though the formal definition of an acknowledgement is fulfilled, the psychological function is not. In such cases it is necessary to suppress the acknowledgement state (essentially reverting to the SR model). It should also be noted that in some cases the only output from the computer consists of the LF and other nonprinting characters. When this situation arises, the output is considered to be response rather than acknowledgement.

2.3.3.1 Parameters Of The SAR Model -

Using the SAR model, several variables characterizing the dialogue can be identified. These variables, identified in Figure 2-2, fall into two broad classes: those concerned with character count, and those concerned with elapsed time. The character count is the number of characters occurring in each message group component.

Each message group component is delimited by two events -- the arrival times of the first and the last characters. Using these two events to measure elapsed time, transmission and delay times are calculated. Transmission time is the time between the first character in a message group component and the last character of that component.

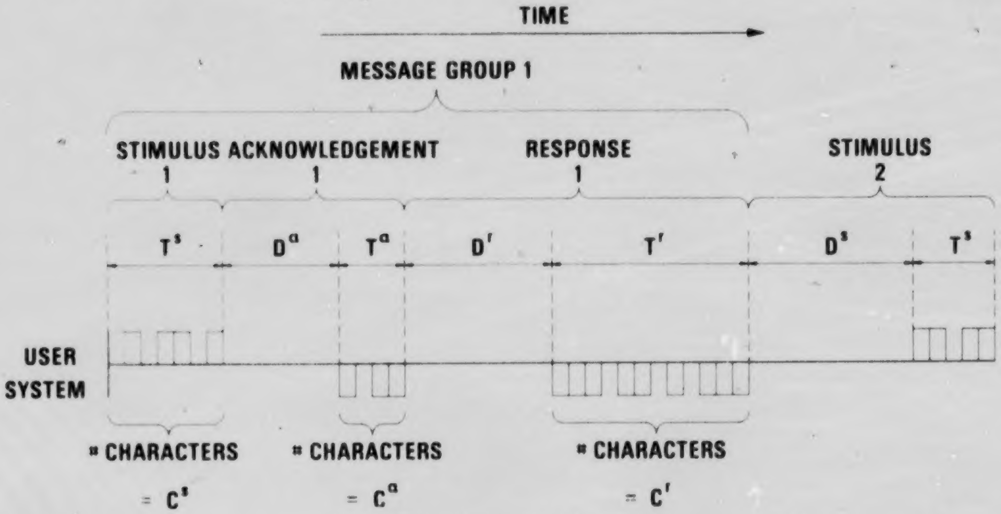


Figure 2-2. SAR Model

The delay time is the elapsed time between the last character of a message group component and the first character of the next component (which may occur in the same or the next message group). Stimulus delay time is the elapsed time between the last character of the network response of the previous message group and the first character of the stimulus of the current message group. Acknowledgment delay time is the interval between the last character of the stimulus and the first character of the acknowledgment, while response delay time is the interval between the last character of the acknowledgment and the first character of the response. The interval between the last character of the stimulus and the first character of the response is the conventional definition of response time.

2.3.3.2 Relationships Among SAR Model Parameters -

Let the following notation be introduced, as illustrated in Figure 2-2:

C^s = Stimulus character count
C^a = Acknowledgement character count
C^r = Response character count

D^s = Stimulus delay time (user think time)
D^a = Acknowledgement delay time
D^r = Response delay time

T^s = Stimulus transmit time
T^a = Acknowledgement transmit time
T^r = Response transmit time

A message group consists of a stimulus-acknowledgement-response 3-tuple. A conversation consists of an ordered set of message groups. Let the subscript i be applied to each of the nine component measures of a message group to indicate the particular message group in a conversation. For example, C_4^s is the stimulus character count for the 4th message group in the conversation.

Analyses may be performed across conversations. Let the additional subscript j refer to the conversation, so that, for example, C_{ij}^s refers to the stimulus character count for the i-th message group in conversation j.

2.3.3.2.1 Compound Delay Times -

Certain other latencies or delay times of interest may be computed from the six basic delay or transmit times.

The stimulus-response delay time (corresponding to the delay time experienced by terminal operators from the completion of the stimulus to the first useful or printing character of the response) may be computed as:

$$D^{sr} = D^a + T^a + D^r \qquad (5)$$

The stimulus inter-arrival time (corresponding to the time from the start of one stimulus to the start of the next, which in some sense may be interpreted as the "duty cycle" for the interactive system) may be computed as:

$$D^{ss} = T_i^s + D_i^a + T_i^a + D_i^r + T_i^r + D_{i+1}^s \qquad (6)$$

That is, the stimulus inter-arrival time is simply the sum of the durations of the stimulus, acknowledgement and response portions of a given message group, or alternatively, the sum of all the transmit times and all the delay times in a given message group.

The total elapsed time for a conversation is the sum of the stimulus inter-arrival times for all the message groups in the conversation. For a conversation with n message groups, the elapsed time is:

$$E = \sum_{i=1}^{n} D_i^{ss} \qquad (7)$$

2.3.3.2.2 Transmission Rates -

The transmission rate of the line or the terminal determines the maximum rate at which characters can be sent or received.

Let

λ = line transmission rate (characters per second)

Then the following burst* transmission rates (measured from the start of the first character in the stimulus, acknowledgement or response) can be computed:

$R^s = C^s / (T^s * \lambda)$ = stimulus transmission rate (user typing speed) (8)

$R^a = C^a / (T^a * \lambda)$ = acknowledgement transmission rate (9)

$R^r = C^r / (T^r * \lambda)$ = response transmission rate (10)

Average transmission rates must include in the computation the delay time prior to the first character. Thus, average transmission rates are less than burst rates due to the idle time when the transmission capacity is unused. Average transmission rates can be computed as follows:

$$\bar{R}^s = C^s / [(D^s + T^s) * \lambda] \qquad (11)$$

$$\bar{R}^a = C^a / [(D^a + T^a) * \lambda] \qquad (12)$$

$$\bar{R}^r = C^r / [(D^r + T^r) * \lambda] \qquad (13)$$

---

* Note that the definition of a "burst" in the SAR model corresponds to the definition of a "burst segment" in the data stream model. In the SAR model, inter-character idle time is considered as part of the burst.

2.4 Interactive Activity Measurement

Empirical research into user characteristics in interactive computer usage, such as the application of the models just discussed, requires a means of collecting data about the user's activity at the terminal, and perhaps about the activities of the computer as well (as least insofar as they affect the user). The basic activity on the part of the user is to enter characters by means of keystrokes; the computer then responds with characters that are either printed or displayed; finally the user reads the response and begins the cycle again by entering more characters.

The user's activities could be measured and recorded directly by a number of techniques popular in experimental psychology. These include observation by a human experimenter with a manual timing device (such as a stop watch) and recording by film or videotape with a timing track or split screen image of the sweep second hand of a watch. Such techniques, however, are cumbersome to use, subject to timing and transcription errors, and generally unsuitable for obtaining a large volume of measurements.

Similarly, the activity of the computer system being exercised could be measured by a variety of internal recording techniques, generally requiring special software running on the same system. This introduces the problems of providing this software for each computer system to be measured, as well as introducing possible pertubations to the very system being measured.

For these reasons, a general, easy to use, automatic recording technique is preferred which can be applied to a large number of subjects without perturbing either the users or the system they are using. This goal can be achieved by focusing the measurement attention on the communications line between the user's terminal and the computer.

When keys are struck by the user, these are electronically encoded into an internal digital representation by the terminal. This sequence of bits is then transmitted serially down the line to the waiting computer where they are assembled back into the original characters. It is relatively easy to construct a passive recording device that may be connected to the same communications circuit (similar to an extension telephone) so that it receives the same signals as the destination terminal or computer. With proper synchronization of the measurement instrument (requiring knowledge of the protocol and transmission rate on the line), these bit sequences may be recognized and recorded as characters. Along with each character can be recorded its time of occurrence (according to some clock in the measurement instrument) and the source of the character (user or computer).

30

Considering the speed with which characters are encoded or printed/displayed by the terminal, the time of detection on the communications line is entirely adequate for the desired purpose, and the timing relationships between *successive* characters is preserved exactly.* Furthermore, such measurements can be made by a passive connection to the line without perturbing either the user or the computer system, and the same method can be used for any computer system employing the same speed/protocol for which the measurement instrument is engineered.

Such a measurement instrument has been designed and constructed by a group of researchers at the National Bureau of Standards to record interactive conversations transmitted according to a start-stop (asynchronous) terminal protocol (Abrams, *et*. *al*., 1977). This instrument and its supporting software has been used in the collection and analysis of data for this dissertation. This data collection and analysis process are the subject of the next chapter.

---

* Note that the propagation delay introduced by certain types of communications systems may influence the choice of *where* the measurement instrument is attached to the circuit.

3.0 DATA COLLECTION

In the previous chapter we discussed the need for empirical measurement of interactive users and systems, developed a model on which to base this measurement, and expressed the need for an automatic, yet passive, measurement instrument. In this chapter we briefly describe the instrument meeting these requirements that was used in the collection of data for this study, the interactive environment in which the instrument was applied, and the type and quantity of data collected.

3.1 Measurement Instrument

The measurement instrument is a combination hardware and software system consisting of a data acquisition device called the Network Measurement Machine (NMM) and a Data Analysis Package (DAP) for generating summary reports. This total system -- the acquisition system and the DAP -- is called the Network Measurement System (NMS). This system was developed at the National Bureau of Standards and is described in detail in a number of reports (Abrams & Cotton, 1975; Rosenthal, Rippy & Wood, 1976; Watkins & Abrams, 1976). We only present enough of a summary here for an understanding of the capabilities and limitations of the system and the type of data collected.

The Network Measurement Machine (NMM) is implemented on a mini-computer employing both regular and special purpose hardware controlled by a specially written software system. The regular hardware includes the processor, an operator's console, disk and magnetic tape storage, two programmable clocks, and data communications interfaces. Special purpose interface hardware is employed to connect the NMM to the interactive system that is to be measured.

Data are not structured or analyzed during acquisition. Rather, all characters are identified, time-tagged and written on magnetic tape with other pertinent information for subsequent analysis. A sequence of interactive sessions for a number of different users can be recorded on the same tape.

Once recorded, the data are processed by the DAP on a large, general-purpose computer system. Briefly, the processing proceeds as follows: The multiple conversations on the tape are first separated into individual conversations. Each conversation is then scanned to build a structure file which contains pointers to the user and computer system messages with their respective time tags according to the SAR model. Different groupings of the data (e.g., by software processor employed by the user) can be noted in the structure file. Analysis of a set of data yields distributions for up to fourteen separate parameters of the SAR model plus line utilization statistics. Conversations may be analyzed individually or in aggregate, reports generated, and a file written for additional data processing by other analysis programs.

3.1.1 Network Measurement Machine -

The purpose of the NMM is to record, characterize, and time-tag selected data dialogues for subsequent analysis. The hardware system required for this specialized data acquisition process consists of a minicomputer (DEC PDP-11/20*) employing both standard and specially designed communications peripherals and related equipment. For detailed information on the NMM, see Rosenthal, Rippy & Wood (1976).

The NMM connects to the communications line between the user's terminal and the computer. Thus, no modifications which might perturb the interactive process are made to the terminal or to the computer system.

In general, data traffic between a terminal and a computer system can be full or half duplex (i.e., two-way alternate or two-way simultaneous). Therefore, it was necessary to provide two separate interfaces, each operating in receive-only mode, to capture the full duplex data dialogue**. Each such pair of interfaces and the associated specialized interconnection hardware for inserting the NMM into a selected data path is collectively called a data probe. Each data probe is "invisible" to the ongoing dialogue with respect to the data, status, and control signals involved. Speed recognition software in the NMM sets the correct operating speed of program-settable (in the range of 110-300 bps) interfaces during initialization.

The acquisition system itself requires software to control and manage the resources of the NMM -- the data probes, the data recording device (magnetic tape), and the operator's console. The NMM software is an interrupt-driven real-time operating system incorporating various device drivers and interrupt service routines for the standard and special purpose peripherals attached to the system.

Regarding the use of the NMM as a data collection device for the present investigation, we may be concerned about the accuracy with which time tags are assigned and thus to which statistics can be computed. In the software, a Data Probe Interrupt Service Routine identifies, time-tags, and buffers communications interrupts. The time-tag clock routine services the crystal controlled clock used to provide 24 bits of timing data. The clock counts at a 10 kHz. rate, providing an interval timer with 100 microsecond resolution. However, the time-tags assigned to characters are only accurate to the nearest millisecond, due to critical non-interruptable code in the service routine (Abrams, et. al., 1977). Thus, when the data are analyzed, statistics can be computed that are accurate to at least the nearest hundredth of a second; in fact, most statistics will only be presented to the nearest tenth of a second, since finer differences do not seem relevent in an environment involving human users.

---

* The identification of certain commercial equipment, including the systems used for data collection and data analysis and the system that was actually measured, is for the purpose of completely describing the performance of the study and does not imply any endorsement on the part of the National Bureau of Standards.

** The capability to record full duplex conversations was built into the NMM as described here. This capability was not required, however, for the application of the NMM in this research (see section 3.2).

3.1.2 Data Analysis Package -

The Data Analysis Package (DAP) processes the data acquired by the NMS. Individual conversations are isolated from the data tape and summary statistics are computed for a number of parameters of interest. The DAP permits data to be grouped in a variety of ways, both within and across conversations. For detailed information on the DAP see Watkins & Abrams (1976).

The DAP is implemented on a Digital Equipment Corporation DECSystem-10. The magnetic tapes recorded by the NMM are transferred to the analysis machine for this processing. The processing optionally produces formatted data files for additional analysis by more sophisticated statistical packages.

Records are ordered on the magnetic tape in the same sequence as they arrive at the NMM. Each NMM-generated magnetic tape may contain data acquired on one or several days of NMM operation; or one day of operation may produce a multi-reel file. In any case, the DAP creates an independent file for each individual conversation represented on the tapes with a numbering convention that reflects the day on which the conversation was recorded. The data in these files are then fit to the analysis model, after which the statistics may be aggregated across sets of conversations.

3.1.2.1 The Analysis Model -

The DAP uses the stimulus-acknowledgment-response (SAR) model discussed in Chapter 2 for structuring the data. The system output must be tested to determine the presence of an acknowledgment. It should be noted that differentiation between acknowledgment and response is semantic and time dependent, that some computer systems issue no acknowledgment, that some systems are inconsistent in their acknowledgment, and that in some cases the acknowledgment constitutes the only response.

Two algorithms are used in combination to determine if an acknowledgement is present. (At the beginning of an analysis session, the analyst using the DAP has the option to specify which of these acknowledgment definitions is to be used or to specify a new definition)*. The first algorithm is based on timing information. If a delay in the network output is encountered greater than a fixed multiple of the character duration, then the output is divided into an acknowledgement and a response. The default parameter is set at 50 character durations; however, the analyst may redefine this parameter.

The second algorithm defines the existence of an acknowledgment based on network output beginning with nonprinting ASCII control characters rather than printing characters. All nonprinting characters occurring at the beginning of network output until the occurrence of a printing character are considered within the acknowledgment (except, as previously noted, that if such non-printing characters are the only system output, they are considered as the response).

The hard copy representation of the conversation is that of the SAR model. The format of the printed record is given in Figure 3-1. Each message group in the conversation is subdivided into its three components, and the characters belonging to each one of these components appear to the right of the corresponding label.

Control characters are optionally represented by their standard abbreviation enclosed in corner brackets. For example, a carriage return would appear as <CR>. Multiple occurrences of control characters are indicated by printing an asterisk followed by the count of repetitions, followed by a closing corner bracket. For example, seven linefeeds would appear as <LF>*7>.

---

* A third algorithm available in the DAP (based on a search for specific character strings) was not used in this study.

```
CONVERSATION RECORD OF FILE 4196#1-1108.HDX;1

<S=STIMULUS, R=RESPONSE, A=ACKNOWLEDGEMENT, E=ECHO>
<SD=STM DLY TIME, ST=STM XMIT TIME, SC=STM CHAR COUNT, SR=STM TRANS RATE>
<AD=ACK DLY TIME, AT=ACK XMIT TIME, AC=ACK CHAR COUNT, AR=ACK TRANS RATE>
<RD=RESP DLY TIME, RT=RESP XMIT TIME, RC=RESP CHAR COUNT, RR=RESP TRANS RATE>
<SRD=STM-RESP DLY, SI=STM INTER ARRIVAL TIME>

S 1      XXXXX
A 1      <CR><LF>*2>
R 1      UNIVAC 1108 TIME/SHARING EXEC <SP>*2>VERS 225 UPDATE B<CR>
         <LF>*2>
                         ST:      22.0    SC:       63    SR:       2.9
         AD:       0.2   AT:       0.2    AC:        3    AR:      17.5
         RD:      .001   RT:       1.8    RC:       51    RR:      28.4
         SRD:      0.4   SI:       0.0

RECORDING TIME: MONDAY, JULY 15, 1974 9:02AM-EDT


S 2      @RUN AAAAA,BBB-CCC,Z<CR>
A 2      <LF><CR>
R 2      DATE: 071574 <SP>*6>TIME: 090315<CR><LF><DEL>
         SD:       4.4   ST:      13.3    SC:       31    SR:       2.3
         AD:      .036   AT:       5.9    AC:        2    AR:       0.3
         RD:       0.2   RT:       1.1    RC:       33    RR:      28.9
         SRD:      6.1   SI:      28.6

S 3      @ED,U ELEMENT.ELT1<CR>
A 3      <LF><CR>
R 3      ED 13.00-07/15-09:03-(27,28) <SP>*2><CR><LF>EDIT <SP>*2><CR>
         <LF>0:
         SD:       0.7   ST:       4.6    SC:       19    SR:       4.1
         AD:      .036   AT:       3.0    AC:        2    AR:       0.7
         RD:       0.2   RT:       1.9    RC:       42    RR:      22.5
         SRD:      3.3   SI:      21.2

S 4      L OUT2<CR>
A 4      <LF><CR>
R 4       <SP>*9>J <SP>*7>OUT2 <SP>*3><CR><LF>101:
         SD:       1.3   ST:       2.2    SC:        7    SR:       3.1
         AD:      .036   AT:       0.5    AC:        2    AR:       3.7
         RD:       0.2   RT:       1.2    RC:       30    RR:      24.1
         SRD:      0.7   SI:      11.0
```

Figure 3-1. Sample Conversation Record

3.1.2.2 Basic Statistics -

The characters with their associated time-tags obtained by the NMM constitute the measured data. The DAP applies the SAR model, resulting in the following basic statistics for each message group (see section 2.2.3.2):

Stimulus character count (C^s)
Acknowledgement character count (C^a)
Response character count (C^r)
Stimulus delay time (D^s)
Acknowledgement delay time (D^a)
Response delay time (D^r)
Stimulus transmit time (T^s)
Acknowledgement transmit time (T^a)
Response transmit time (T^r)

In addition, the following parameters are computed, based on the arithmetic relationships given in the indicated equations in Chapter 2:

Stimulus-Response delay time (D^{sr}) (5)
Stimulus inter-arrival time (D^{ss}) (6)
Stimulus transmission rate (R^s) (8)
Acknowledgement transmission rate (R^a) (9)
Response transmission rate (R^r) (10)

The intent of this research is the measurement of activity typical of a user/system dialogue, not the measurement of anomalies; therefore, it is reasonable to eliminate outliers for the calculation of the statistics. For example, it is possible for an on-line network user to become distracted by and involved in an activity totally unrelated to network usage. It is also possible for an interactive system to "crash" at any point during a conversation. Such events produce distorting data. To recognize the presence of these data, upper and lower limits are used. These limits determine the standard sampling interval. Data must fall within the interval to be considered in the statistics computed by the DAP.* The lower limit for all parameters is set at zero. Table 3-1 contains the upper limits for the nine basic and four derived parameters which, together with the lower limit of zero, defines the standard sampling intervals for these parameters.

---

* The file written for processing by other statistical routines contains all the data, including outliers.

Parameter	Unit	Upper Limit
Stimulus character count	characters	60
Acknowledgement character count	characters	60
Response character count	characters	300
Stimulus delay time	seconds	60
Acknowledgement delay time	seconds	20
Response delay time	seconds	20
Stimulus transmit time	seconds	50
Acknowledgement transmit time	seconds	15
Response transmit time	seconds	45
Stimulus inter-arrival time	seconds	300
Stimulus transmission rate	cps	10, 15 or 30
Acknowledgement transmission rate	cps	10, 15 or 30
Response transmission rate	cps	10, 15 or 30

Table 3-1. Parameter Upper Limits.

By dividing the standard sampling interval into a number of subintervals it is possible to characterize the distribution of the derived parameters by counting the number of occurrences of a parameter value in each of the subintervals. In addition to the histograms, statistical measures of the data including the mean, standard deviation, median (50th percentile), and the 90th and 95th percentiles are computed.

The printed output for each conversation analyzed can optionally include any of the following: statistics for selected SAR model parameters, histograms for these parameters, a conversation summary and line utilization statistics. The summary begins with a review of the statistics associated with each selected parameter. The speed of the connection (recorded in the configuration record), the number of occurrences of anomalous data (values occurring outside the standard sampling interval), and the total time of the conversation are printed. The line utilization statistics summarize the activity on the communication line over which the conversation took place. All characters are generated by either the user or the network; further, they are either printing or nonprinting. A variety of percentages relative to these character groupings are calculated which serve as a profile describing usage of a connection.

Communications channel utilization is represented as the percentage of actual use relative to the potential use. Two measures of utilization are given. One defines the potential time interval as beginning with the first character of a message sent by the source and ending with the last character of that message. The other measure incorporates in its calculation of the potential time interval the delay time imposed by the source These statistics help to indicate if the user has chosen an unrealistic connection speed.

A summary file which contains the frequency distribution array for the standard sampling interval obtained from the analysis is created.* The analysis of multiple conversations is performed by creating a composite frequency distribution by totaling the contents of the grouped frequency counts in the summary files selected. Such analysis results in the same type of parameter statistics, histograms and conversation summaries that are produced for individual conversations.

3.1.2.3 Subsets -

The working unit in the interaction between the user and the system is the message group, which may be considered a generalization of a transaction. Employing set terminology, the conversation is the ordered set of all message groups from logon to logoff. Many other sets of message groups can be defined. The set concept may also be extended beyond the boundary of a single conversation, where the upper limit is the entire data base obtained by the NMM. For example, when all of the usage of a given network during a period of time such as a month is considered, the set encompasses multiple conversations.

In the present context the interest is in sets which encompass less than a conversation. Message groups may be identified according to the functional objective with which they are associated. For example, in a programming environment the use of the various language translators, the editor, the linking loader, and the execution of debugging tools could each constitute a subset. There is no requirement that the subset definitions be mutually exclusive.

Subset identification makes it possible to take various samplings or cross sections through the data base, depending upon the objective of the analysis. Using the editor as an example, it is possible to identify what percentage of the message groups or the elapsed time is spent in the use of this resource. It is also possible to limit the attention to this editing resource and to perform all of our statistical analysis on it. Subsets may be identified manually or by special programs after the structure file has been built according to the SAR model. Assigning a message group to a subset is accomplished by setting a bit in a "subset mask word" in the header for each message group.

The statistical routines described earlier are available for operation on the subsets. They may be applied to individual subsets or to logical combinations of subsets. When subsets are analyzed, statistics are aggregated for any message groups designated by the proper bits in the mask word. Summary files

---

* This is not the file optionally produced for processing by other analysis routines.

(containing the frequency distribution array) are created for each subset (or logical combination of subsets) so that subsets may be aggregated across conversations. Files containing an array of all the individual parameter values for each qualifying message group may also be created for subsequent analysis by other programs.

3.1.3 Additional Processing -

Whenever the DAP are used to analyze a set of message groups, either within a single conversation or across conversations (totally or by subset), all the values for the 14 parameters for each message group can optionally be written into a file. This capability was not included in early versions of the DAP, but was requested by the author for additional flexibility in analysis. These files are very simply formatted, and may be viewed as data arrays of 14 x N, where N is the number of message groups for which values have been computed. Once the data have been written into these files, there is no way to relate the data back to the specific conversation where it originated.

Several routines were written by the author to manipulate these data files. One routine simply computes all the cumulative percentiles from 5% to 95% at 5% intervals, for selected parameters. Another routine permits data to be trimmed at specified high or low values, or both, prior to these percentiles being computed. This routine permits all the data for a message group to be discarded based on trimming any single parameter.

One routine was written to compute serial or autocorrelation coefficients (See section 4.1.2) for all the parameters of a set of data. This routine also wrote these values into a new file in the same format. This new file (of coefficients computed for sets of data) could then itself be analyzed to determine the relative distribution of coefficients.

Finally, two routines were written which served as interfaces to library programs for the Mann-Whitney and Kolomogorov-Smirnov statistical tests (see section 4.2). These interface routines passed data to the library programs in the proper way to compare two distributions of a single parameter, and printed the test results.

All of these programs were written in FORTRAN and run on the DECSystem 10 computer in the Institute for Computer Sciences and Technology of the National Bureau of Standards. They have been used extensively to process the data collected in this study and to prepare the results presented in the following chapters.

3.2 Interactive Environment

All the data collected in this study were on interactive users of the central Univac 1108 installation at the National Bureau of Standards in Gaithersburg, Maryland. This system supports both the scientific and administrative data processing needs of the Bureau. Interactive terminals were supported on the system only in half-duplex mode, at speeds in the range 110-150-300 bits per second (bps).* Prior to 1974, line speeds were pre-set for different dial-in lines; in mid-1974 a front-end processor was installed that provided speed recognition from among these three speeds on the same dial-in line.

In order to collect data on users of this system, an extension telephone line connected to a dial-in port of the Univac 1108 was installed at the site of the NMM (in another building). This extension is terminated by two modems operating in receive-only mode, wired to never disconnect (hang-up). One modem is strapped to always demodulate the high band, the other to always demodulate the low band.** Data is accepted by the NMM only when the carrier is present in both modems. The conversation is assumed to begin when carrier is detected and the speed recognized by the NMM, and to end when carrier is lost.

The terminal population at the National Bureau of Standards includes a wide variety of different terminals, including CRTs and printing terminals, from old model teletypewriters to the newest CRT terminals. Unfortunately, there is no way to determine which type of terminal was used for any particular conversation, though there are some generalizations that can be made. Conversations recorded at 10 characters per second were unlikely to have occurred on other than model 33 or older teletypewriters, since newer model teletypewriters and CRTs can operate at higher speeds, and since users are unlikely to operate their terminal at less than the maximum supported rate. Conversations at 15 cps were most likely made with model 37 teletypewriters or General Electric Terminet terminals, since this is the maximum rate for these devices. Conversations at 30 cps could have been produced on any of a wide variety of terminals, but obviously not on any of the terminals whose maximum speed is less than 30 cps.

---

* Equivalent to 10-15-30 characters per second (cps).

** As implemented in Bell 103 compatible modems, Frequency Shift Keying (FSK) employs two frequency bands for data transmission. The low band frequency is modulated by the call originator (modem of the source device) and the high band is modulated by the call responder (modem of the destination device).

41

3.3 Measurements Taken

With the use of the remote telephone extension described above, interactive users on the NBS Univac 1108 were recorded on randomly selected days in the three year period of 1974-1976. Users were informed by a note in the regular installation newsletter that a particular dial-in number was subject to being monitored. Thus, users with sensitive applications or who, for any reason, did not wish to be recorded, could have avoided the use of this particular port. Other than this general notice, there was no specific indication when recording started or ended, and dial-in users on the line subject to recording had no way of knowing whether they were being recorded or not. Since the recordings and associated conversation listings include user passwords, they have been guarded carefully.

Recording generally took place for several hours at randomly selected times on randomly selected days. No specific technique was employed for selecting these days; rather, recordings were taken whenever the equipment was functioning and personnel were available to operate it. In this manner, data were successfully recorded on the following number of days in each of the three years covered:

1974 - 31 days
1975 - 39 days
1976 - 35 days

Total - 105 days

Table 3-2. Summary of Data Recording

3.3.1 Selection Of Conversations -

All the data collected on the 105 days of recording were processed by the Data Analysis Programs to isolate the individual conversations, fit the data to the SAR model, assign subsets, and generate summary statistics for the nine fundamental and five derived parameters. The conversations were carefully culled to remove any which did not represent normal interactive use of the NBS 1108 installation. Thus, conversations with other systems recorded on the same tape, conversations where the speed had been improperly recognized or where numerous parity or other types of errors occurred on the line between the user and the computer, and conversations where interactive use was known to have been interrupted for paper tape input, were eliminated from consideration. Also, conversations that were less than several minutes in duration or which did not include at least 10 message groups were eliminated since they were not representative of normal interactive sessions.* The ratio of acceptable conversations to total conversations recorded was remarkably low -- only about one in three conversations was deemed suitable for further processing. Many conversations were discarded because the user failed to interact properly with the system, e.g., short conversations consisting entirely of improper and thus unsuccessful login attempts.**

The result of this culling was the following number of "good" and "representative" conversations for each of the calendar years indicated:

1974 - 78 conversations
1975 - 119 conversations
1976 - 86 conversations
Total - 283 conversations

Table 3-3. Tabulation of "Good" Conversations by Year

Following the processing to fit the conversation data to the SAR model, the speed of the conversation and the duration, both in terms of time and number of message groups, was available. Table 3-4 presents the results of data collection, tabulated by terminal speed.

Grouping by terminal speed was accomplished by aggregating the statistics for entire conversations identified as occuring at the same speed. The results of analyzing the data grouped in this way are discussed in Chapter 5.

---

* The lower limit of 10 message groups to an acceptable conversation is admittedly somewhat arbitrary, but was chosen after examination of a large number of brief conversations revealed most to be representative of error conditions rather than normal operation.

** Data on such unsuccessful login attempts may be interesting and useful in the study of the ease of interactive system use. However, this was not the object of the present study so these data were not tabulated.

	Terminal Speed (bits per second)		
	110	150	300
Number of conversations			
1974:	-	-	78
1975:	23	5	91
1976:	10	2	74
Total	33	7	243
Number of message groups:	2638	361	19706
Total time (minutes):	964.2	139.6	5364.8

Table 3-4. Tabulation of "Good" Conversations by Speed

3.3.2 Subset Assignment -

The nature of subsets was discussed in Section 3.1.2.3. For the Univac 1108, subsets may be assigned automatically by a program which scans for the "master space" character (@) that begins all executive level control statements. Such a program was written by a research assistant under the author's direction. A different subset is defined for each distinctive class of executive level statement, and that statement and all subsequent message groups up to the next recognized executive level statement are assigned to the subset.

The grouping by application is accomplished by aggregating statistics for all message groups assigned to the same subset. Due to the relative paucity of data at other speeds, only the 300 bps data was assigned to subsets and so analyzed. The results of this grouping and analysis is discussed in Chapter 6.

4.0 DATA ANALYSIS

In Chapter 3 we outlined the experimental environment, the data collection process, the quantities of data collected and the basic data analysis that was considered part of the data collection process (viz., the application of the SAR model). In this chapter, we discuss the statistical treatment of the data collected, including the experimental design for data analysis and tests of significance. In succeeding chapters we will discuss the results of the application of these data analysis procedures for a number of different experimental designs.

4.1 Review Of Data Collected

As explained in the previous two chapters, the basic events recorded by the measurement instrument are the source, code and time tag for each character transmitted by either the user or the system. Groups of characters are fit to the SAR model, resulting in a series of stimulus-acknowledgement-response triplets for each conversation. The statistics available for analysis are in terms of message groups, not individual characters. Fourteen parameters are computed for each message group as summarized in Table 4-1. Thus, for each conversation, the available data consists of a matrix of dimension 14 x N, where N is the number of message groups in that conversation.

Parameter	Type	Description
D^s	Basic	Response-stimulus delay (think) time
T^s	Basic	Stimulus transmit time
C^s	Basic	Stimulus character count
R^s	Derived	Stimulus transmission rate
D^a	Basic	Stimulus-acknowledgement delay time
T^a	Basic	Acknowledgement transmit time
C^a	Basic	Acknowledgement character count
R^a	Derived	Acknowledgement transmission rate
D^r	Basic	Acknowledgement-response delay time
T^r	Basic	Response transmit time
C^r	Basic	Response character count
R^r	Derived	Response transmission rate
D^{sr}	Derived	Stimulus-response delay time
D^{ss}	Derived	Stimulus inter-arrival time

Table 4-1. Summary of Message Group Parameters

4.1.1 Possible Groupings -

Compound analyses may be performed by aggregating the data for all the message groups in a single or group of conversations, or by aggregating selected message groups in a single or group of conversations. These alternatives are summarized below in Table

4-2, with a suggestion of the most suitable research question for each.

	MESSAGE GROUPS	
CONVERSATIONS	All	Selected
Single	Behavior of one user	Behavior of one user on selected tasks
Multiple	Typical behavior of all users (interactive workload)	Typical behavior of all users on selected tasks

Table 4-2. Data Analysis Alternatives

The behavior of individual users is not the concern of this study; hence, the analysis of single conversations, either entirely or in part, has not been undertaken. The intent of this study is to investigate the typical behavior of groups of users in general and on specific tasks, and the performance of the system to such typical demands. Thus, the types of analyses to be described in subsequent chapters fall into one of the bottom two cells in Table 4-2.

4.1.2 Data Independence -

One question that needs to be addressed whenever data are batched is whether the individual observations are independent of one another. This question is particularly important in the case of a time series of observations, such as we have in the sequence of message groups comprising an individual conversation. If the observations prove to be highly correlated in a serial fashion, then we will not be justified in aggregating them as individual observations with data from other conversations.

The question of independence in a time series can be investigated by means of the autocorrelation or lag-1 correlation coefficient which provides a measure of the overall correlation between each successive pair of observations. This statistic is computed as follows (for a series of N observations):

$$\rho_1 = \sum_{i=1}^{N-1} \frac{[(x_i - X)(x_{i+1} - X)]}{s_i^2(N-1)} \tag{14}$$

where X is the mean of the x_i
and s_i is the standard deviation of the x_i

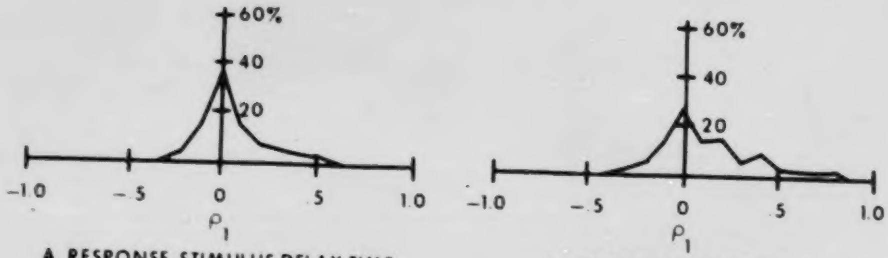
Mamrak and DeRuyter (1977) suggest that "a reasonable rule of thumb for deciding if the data are independent" is to take $|\rho_1| < 0.1$ as evidence of no correlation. They further suggest batching of data where volume permits (e.g., taking the mean of every n data items as the elementary datum) as a way of eliminating autocorrelation.

Statistical techniques exist for determining the probable degree of autocorrelation within a single sample such as the successive message groups in a conversation. However, in a large set of samples when, for example, the autocorrelation coefficient is tested to be less than 0.1 to some confidence limit, we would still expect some samples to fail the test purely due to random variation. If the degree of confidence for the test were chosen to be 95%, we would expect 5 out of every 100 samples drawn from a truly uncorelated underlying population to fail the test. Considering that the number of samples in this study is in the hundreds, a procedure is required to assess the degree of autocorrelation for each parameter over the complete set of samples.

One way to do this is to consider the distribution of autocorrelation coefficients for all conversations. A computer program was written under the author's direction to compute ρ_1 for all fourteen parameters for all conversations. If the data for each parameter are not serially correlated, we might expect ρ_1 to be normally distributed around a mean of zero. To determine this, the values of ρ_1 for each parameter were plotted in relative frequency histograms (Figures 4-1-a through 4-1-n). Even without formal tests of central tendency, it can be readily seen that most system-related parameters do indeed have zero as the central tendency for ρ_1, while the user-related (stimulus) parameters have some moderate degree of positive autocorrelation.

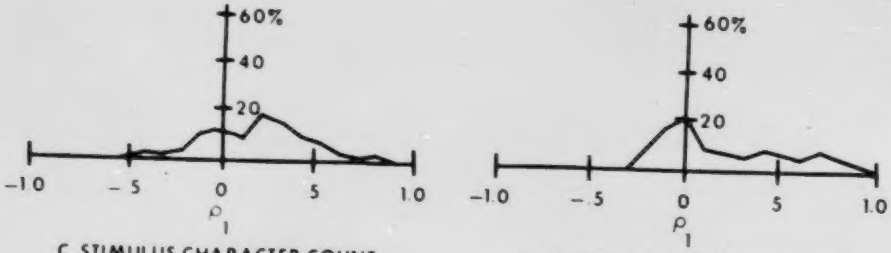
For those parameters where autocorrelation is not evident, all the data collected can safely be used. However, for parameters where autocorrelation is evident, use of all the data could yield misleading results, since all the observations would not be truly independent. For these cases, a compression of the data is required to obtain independent observations.

One method suggested by Mamrak and DeRuyter (1977) for eliminating autocorrelation is to batch the data (e.g., take the mean of every n data items as the elementary datum). They observe that the batch size (n) can be increased (where volume permits) until the serial correlation is reduced to an acceptable level.



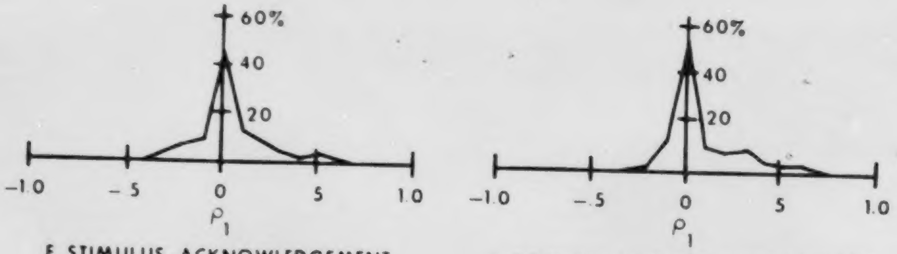
A. RESPONSE-STIMULUS DELAY TIME

B. STIMULUS TRANSMIT TIME



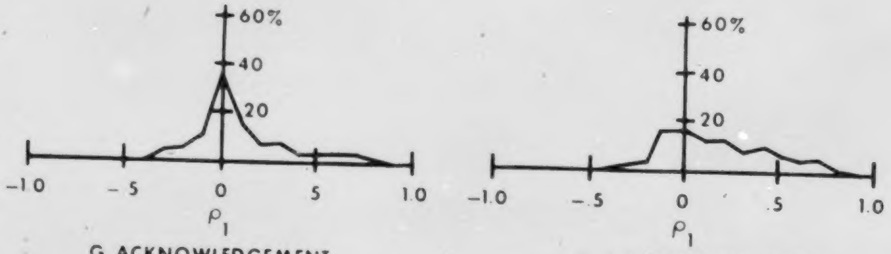
C. STIMULUS CHARACTER COUNT

D. STIMULUS TRANSMISSION RATE



E. STIMULUS-ACKNOWLEDGEMENT DELAY TIME

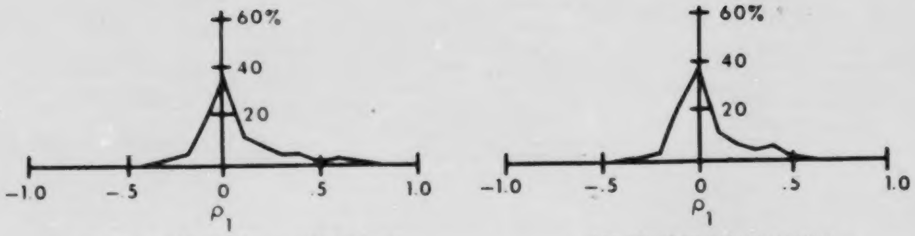
F. ACKNOWLEDGEMENT TRANSMIT TIME



G. ACKNOWLEDGEMENT CHARACTER COUNT

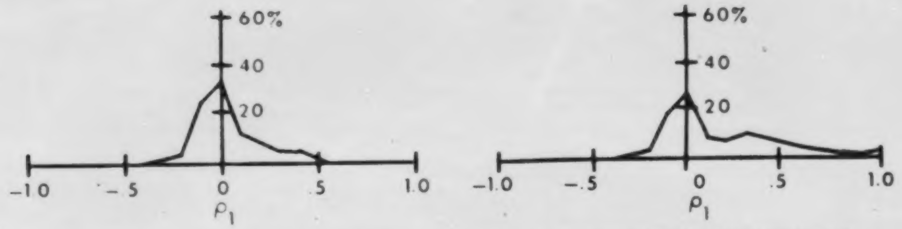
H. ACKNOWLEDGEMENT TRANSMISSION RATE

Figure 4-1. Distribution of Autocorrelation Coefficients of Each Conversation, by Parameter



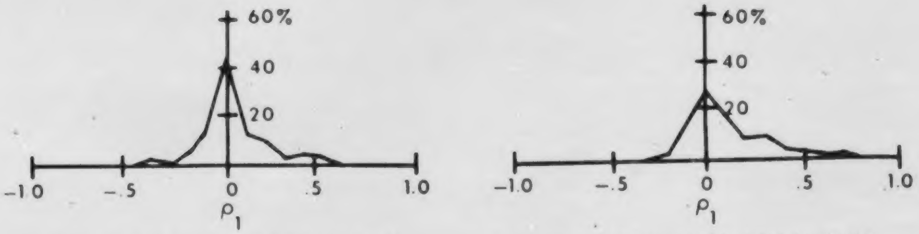
I. ACKNOWLEDGEMENT-RESPONSE DELAY TIME

J. RESPONSE TRANSMIT TIME



K. RESPONSE CHARACTER COUNT

L. RESPONSE TRANSMISSION RATE



M. STIMULUS-RESPONSE DELAY TIME

N. STIMULUS INTER-ARRIVAL TIME

Figure 4-1. (Continued)

For the data collected in this study, it is reasonable to expect serial correlation between the message groups in each individual conversation, but not across conversations. For this reason, each conversation represents a convenient batch for the set of message groups within it. The mean is still an inappropriate choice of test statistic, since it is so highly influenced by outliers, so that the median can be chosen again as a better test statistic. Thus, by forming the distributions of parameter medians for each conversation in a terminal speed class, the same statistical tests used for all data can be applied without violating the assumption of independent observations. However, the sample size is greatly reduced when this is done, so that the significance levels of the tests will not be as great as with all the observations used.

4.2 Non-Parametric Tests Of Significance

Inspection of the distributions obtained empirically for each of the fourteen parameters reveals that none of them appears to be distributed normally. Thus, the parametric tests of significance that are ordinarily applied may not be used directly. It would be possible to seek a transformation that would result in a normal distribution; however, if tests can be found that work adequately without the necessity for such a transformation, they would certainly be preferable. Two candidate tests are the Mann-Whitney U-test and the Kolmogorov-Smirnov test, neither of which depend on any assumptions about the shape of the underlying distribution (i.e., they are non-parametric tests).

4.2.1 Mann-Whitney U-test -

When at least ordinal measurement has been achieved, the Mann-Whitney U test may be used to test whether two independent groups have been drawn from the same population. According to Siegel (1956), this is one of the most powerful of the non-parametric tests, and "it is a most useful alternative to the parametric t-test when the researcher wishes to avoid the t-test's assumptions..."

The Mann-Whitney U-test is quite simple in concept and easy to perform. The null hypothesis is that, given samples from two populations, the populations have the same distribution. Alternate hypotheses can either be one-tailed (directional) or two-tailed. To test the hypothesis, the samples are combined and ranked in increasing order. The test statistic U is given by the total number of times that a score in the sample group with fewer observations preceeds a score in the group with the larger number of observations. The contribution of Mann and Whitney (1947) was to show that for large sample sizes (>20) the distribution of U rapidly approaches the normal distribution, with

$$\text{Mean} = \mu_u = \frac{n_1 n_2}{2} \qquad (15)$$

and $$\text{Standard deviation} = \sigma_u = \frac{(n_1)(n_2)(n_1+n_2+1)}{12} \qquad (16)$$

Thus, for $n_i > 20$ the significance of an observed value of U may be determined by

$$z = (U-\mu_u)/\sigma_u \qquad (17)$$

which is practically normally distributed with zero mean and unit variance.

A subroutine for performing the Mann-Whitney U-test is provided in the IBM Scientific Subroutine Package (IBM, 1968). A program was written by the author to permit selected data collected by the Network Measurement System to be tested by the Mann-Whitney procedure using the subroutine. The values of U, z and the associated probability are printed whenever the program is run for a set of data. This program was used in the analyses to be described in Chapter 5.

4.2.2 Kolmogorov-Smirnov Test -

The Kolmogorov-Smirnov two-sample test is another test of whether two independent samples have been drawn from the same population (or from populations with the same distribution) (Smirnov, 1948). The two-tailed test is sensitive to any type of difference in the distributions from which the two samples were drawn, e.g., differences in central tendency, dispersion, higher moments. The one-tailed test is used to decide whether or not the values of the population from which one of the samples was drawn are stochastically larger than the values of the population from which the other sample was drawn.

The actual test is concerned with the agreement between two cumulative distributions. If the two samples have been drawn from the same population distribution, then the cumulative distributions of both samples may be expected to be fairly close to each other, since they both should show only random deviations from the population distribution. If the two sample cumulative distributions are "too far apart" (Siegel, 1956) at any point, this indicates that the samples likely came from different populations. Thus, a large enough deviation between the two sample cumulative distributions is evidence for rejecting H_o.

To apply the Kolmogorov-Smirnov two-sample test, a cumulative frequency distribution is constructed for each sample, using the same intervals for both distributions. Then, the difference between the two distributions is computed for each interval. The test focuses on the largest of these differences.

Let $S_{n_1}(X)$ = the observed cumulative step function of one of the samples, that is, $S_{n_1}(X) = K/n_1$, where K = the number of scores equal to or less than X. Similarly for S_{n_2}. Then the Kolmogorov-Smirnov two-sample test focuses on

$$D = \text{maximum } |S_{n_1}(X) - S_{n_2}(X)| \qquad (18)$$

The statistic $\sqrt{\frac{n_1 n_2}{n_1+n_2}}\, D_{n_1,n_2}$ is a random variable with limiting cumulative distribution function L(z) as follows:

$$L(z) = \lim_{n_1,n_2 \to \infty,\infty} \text{Prob}\left[\sqrt{\frac{n_1 n_2}{n_1+n_2}}\, D_{n_1,n_2} < z\right] \qquad (19)$$

The probability (asymptotic) of the statistic $\sqrt{\frac{n_1 n_2}{n_1+n_2}}\, D_{n_1,n_2}$ being not less than its computed value, under the assumption (null hypothesis) of equality of the two theoretical distribution functions from which the two samples were taken, is computed

$$P = 1 - L(z) \qquad (20)$$

A subroutine to perform the Kolmogorov-Smirnov test is also provided in the IBM Scientific Subroutine Package, and the author prepared a program to utilize this subroutine for the analysis of sets of data collected by the Network Measurement System. The values of z and its associated probability are printed whenever the program is run for a set of data. This program was also used in the analyses to be described in Chapter 5.

4.2.3 Ranking And Selection Techniques -

When the intent of statistical analysis is to rank the performance of some number of different classes of observations (of some number of interactive systems) in order to be able to select the "best" with some level of confidence, then a variety of different techniques can be used, based on sample means, percentiles, or proportions. These techniques are based principally on the work of Sobel (1967); general descriptions of the procedures and their applicability have recently become available (Gibbons, Olkin and Sobel, 1977). When applied to selection problems, the procedures specify the number of data points which must be collected from each alternative in order to guarantee that the probability of a correct selection is as least as great as some predetermined value.

These techniques have recently been applied to computer comparison studies, using data collected with the same measurement tool described in the previous chapter (Mamrak and DeRuyter, 1977; Amer and Mamrak, 1978; Mamrak and Amer, 1978). In a real sense, this work has been going on in parallel with the present study (and with some contact between the researchers). However, the purpose of these studies was much more specific (aimed at procedures for computer selection) than the current study (aimed at investigating the relevent factors in interactive computing in general). Thus, these procedures were not appropriate for use in the current study, though they are available and now sufficiently well described (and with the needed tables) to be used where required.

5.0 EFFECTS OF DIFFERENCES IN TERMINAL SPEED

Previous chapters have dealt with the motivation for this research, the methodology for data collection and the data analysis procedures. In this chapter we discuss the results of analyzing the data according to terminal speed. This serves to provide one example of the methodology as well as to obtain some practical results, viz., descriptive data about the performance of users (at terminal rates higher than most studies previously reported) and the performance of a representative interactive system.

As was explained previously, data were collected over a three year period for interactive users of the Univac 1108 system at the National Bureau of Standards. Due to the availability of a wide variety of terminals during that time period, data were collected for users operating terminals at 110, 150 and 300 bits per second (10, 15 and 30 characters per second). The data collected were aggregated according to terminal transmission rate and analyzed to isolate the effects of varying the terminal rate on user performance and on system performance (as evidenced by the parameters of the SAR moded).

5.1 Experimental Design

The basic experimental design for this chapter is to compare the effects of different treatments (terminal speed) according to a number of parameters (the 14 parameters of the SAR model). Actually, the distribution of parameter values for each treatment group will be compared for each parameter; the null hypothesis, H_0, is that there is no difference in the distributions, i. e., that the samples were actually drawn from the same population. The non-parametric tests discussed in Chapter 4 will be used to test the significance of observed differences in the distributions. Although there were actually three different terminal speeds for which observations were collected, so few data were collected at the 150 bps rate that these data are not tested. (Many of the results for the 150 bps data appear to be anomolous due to the relatively low number of observations). Thus, only a pairwise test is performed comparing each of the treatments (110 bps versus 300 bps) on each of the 14 parameters, using both the Mann-Whitney U-Test and the Kolmogorov-Smirnov Test. These tests are applied both for the distributions of all the observations and for the distributions of conversation medians, as was discussed in section 4.1.2.

All of the usable data collected in the study are analyzed in this chapter. A summary of the number of "good" conversations recorded at each data rate and the total number of message groups in each such set of conversations may be found in Table 3-3. However, there are not as many data items for each parameter as the number of message groups would indicate. Message groups in each conversation (particularly those which initiate or terminate the conversation) may be missing certain portions (stimulus, acknowledgement or response). Table 5-1 summarizes the number of observations that were found for each parameter for each data rate when the data were analyzed.

	110	150	300
D^s	1471	351	17763
T^s	2638	356	19639
C^s	2638	356	19705
R^s	2638	356	19705
D^a	1362	323	17993
T^a	1601	330	18330
C^a	1489	330	17545
R^a	1601	330	18330
D^r	1334	184	9741
T^r	2553	331	17683
C^r	2537	331	17501
R^r	2553	331	17683
D^{sr}	1578	319	16907
D^{ss}	2615	349	19444

Table 5-1. Number of Message Groups by Parameter and Data Rate

In the following sections, the empirically observed values for each of these fourteen parameters are presented, both in graphical and abbreviated tabular form, as cumulative frequency distributions. It is felt that this is the most meaningful way in which to present and interpret the results, since the ordinary frequency histogram for all of these parameters is decidedly non-normal. Thus, as just discussed in Chapter 4, non-parametric tests are used to compare differences between distributions. These tests themselves are based on the use of cumulative distributions. Presenting cumulative information also minimizes the effect of outliers, either at the low or high end of the distribution function.

When these parameters are used in practice (e.g., when specifying design goals), certain points on the distribution (equating to certain cumulative percentiles) are typically specified. The most commonly specified of these, the median or 50 percentile and the 90 percentile, are tabulated separately from the complete graphical presentation.

An adequate number of observations has been obtained to statistically compare the distributions of parameters for terminals at the 110 bps and 300 bps rates. The number of observations for the 150 bps terminals does not appear to be adequate; hence, these data are not analyzed and are presented for completeness only. Many of these distributions appear to be anomolous.

The statistical significance of the observed differences between the distributions of the model parameters in each case is an important question. For this reason, though the tests results are presented for each parameter in the following sections, the issue of statistical significance is discussed in a separate section (5.5) after the empirical results have been discussed. To summarize the test results, when the Mann-Whitney U-test and Kolmogorov-Smirnov test are applied to compare the distributions of all parameters for the 110 bps case and the 300 bps case, using all the data in each case, all the differences between distributions are significant at the 1% level or better. However, a degree of serial correlation was noted for some parameters in section 4.1.2. When the data are batched so that a single statistic is used for each conversation (the median of each parameter) in order to eliminate the serial correlation, the differences between the distributions for a few parameters are not significant at even the 5% level.

5.2 Effects Of Terminal Speed On User Performance Parameters

The user performance parameters are those associated with the stimulus portion of the message group. These parameters include response-stimulus delay time (also called "think" time), stimulus transmit time, stimulus character count and stimulus transmission rate. What is of interest here is to ascertain the effect on the user (changes in user performance as evidenced by the effect on the four parameters descriptive of the stimulus portion of the message group) of an increase in the operating speed of the terminal.

5.2.1 Response-Stimulus Delay (Think) Time -

The Response-Stimulus delay time (Ds) is the time in seconds between the last character in the response portion of message group N and the first character in the stimulus portion of message group N+1. It is characterized as the "think" time, since it represents the latency period during which the user "digests" the response to the previous stimulus and formulates the next stimulus.

Measurements of Response-Stimulus delay time for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-1. The key percentile values for each class are tabulated in Table 5-2, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

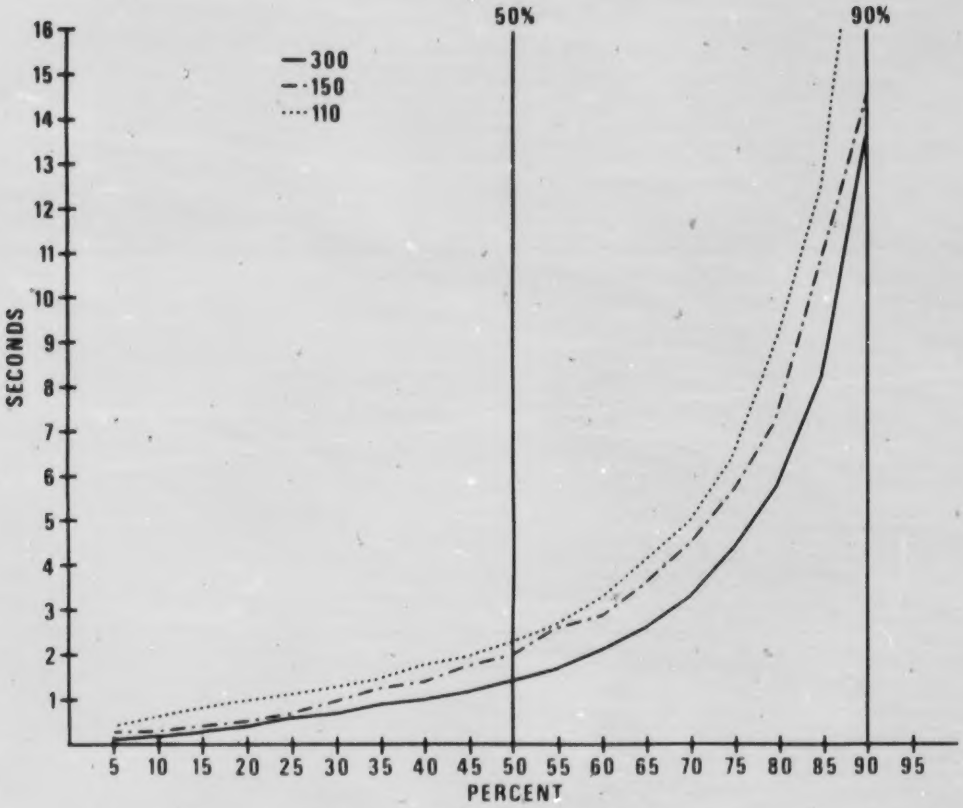


Figure 5-1. Cumulative Frequency Distribution of Response-Stimulus Delay Time

	All Observations*		Conversation Medians**	
Rate	50%	90%	50%	90%
110 BPS	2.3	19.7	2.3	6.1
150 BPS	2.0	14.6	1.9	4.8
300 BPS	1.4	13.6	1.5	4.2

Table 5-2. Key Percentiles for Response-Stimulus Delay Time (Seconds)

These data show a clear reduction in user think time as the operating speed of the terminal is increased. Since the interval for user think time does not begin until the last character of the response (from the previous message group) has been sent, this reduction can not be attributed to the ability of the operator to read ahead while the response is being printed. If anything, reading ahead would tend to reduce the think time on lower speed terminals, since the text of the response is displayed for a longer period of time before the think time interval begins. Thus, the reduction in user think time must be attributed to a general increase in the pace of operator actions elicited by the opportunity to interact with the computer at a higher rate.

---

* Number of observations as given in Table 5-1.

** Number of observations as given in Table 2-3.

Note: * and ** apply for all successive tables of percentiles.

5.2.2 Stimulus Transmit Time -

The stimulus transmit time (T^s) is the time in seconds from the first to the last character of the stimulus. This parameter refelcts both the user rate of data entry and the number of characters entered. By itself, this parameter is an indication of the transmission time required by the operator on the channel from the terminal to the computer system. The stimulus character count (parameter 3) is divided by the stimulus transmit time to compute the operator input or typing rate.

Measurements of stimulus transmit time for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-2. The key percentile values for each class are tabulated in Table 5-3, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	0.6	13.9	3.7	6.8
150 BPS	2.9	11.3	2.4	3.5
300 BPS	2.0	13.8	2.5	6.1

Table 5-3. Key Percentiles for Stimulus Transmit Time (Seconds)

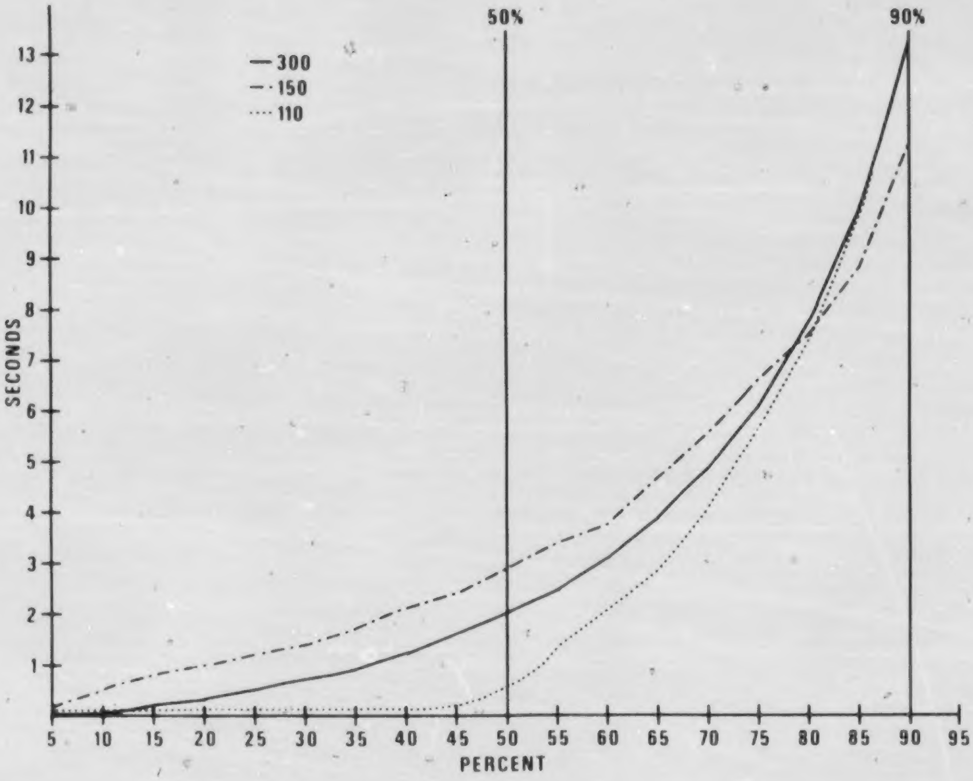


Figure 5-2. Cumulative Frequency Distribution of Stimulus Transmit Time

Stimulus transmit time shows an increase in the median value for all data as the terminal speed is increased from 110 to 300 bps (the value for 150 bps is interpreted as being anamolous). The 90 percentile values are about the same. However, this parameter exhibited a certain degree of serial correlation (as evident in Figure 4-1-b). When the serial correlation is eliminated by taking the distributions of conversation medians, a different picture emerges: stimulus transmit time is seen to decrease as the terminal speed increases.

The change in stimulus transmit time as a function of terminal speed represents the net effect of two separate changes working in opposite directions. Any increase in stimulus character count with increased terminal speed would tend to increase the stimulus transmit time, while any increase in stimulus transmission rate would tend to decrease stimulus transmit time. Both of these tendencies are evident to some degree (see the following two sections); apparently, the increase in stimulus transmission rate more than makes up for any increase in stimulus character count.

5.2.3 Stimulus Character Count -

The stimulus character count (Cs) is the number of characters entered by the operator as input to the computer system.

Measurements of stimulus character count for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-3. The key percentile values for each class are tabulated in Table 5-4, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level without batching of the data; however they are not significant at an acceptable level when the data are batched (see section 5.5).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	2	21	7	14
150 BPS	7	21	6	8
300 BPS	6	23	7	13

Table 5-4. Key Percentiles for Stimulus Character Count

The increase in stimulus character count for all observations that is associated with higher speed terminals is one of the more interesting results from the measurement of user performance parameters. Apparently, some users of higher speed terminals feel freer to express themselves somewhat more verbosely to the computer system.

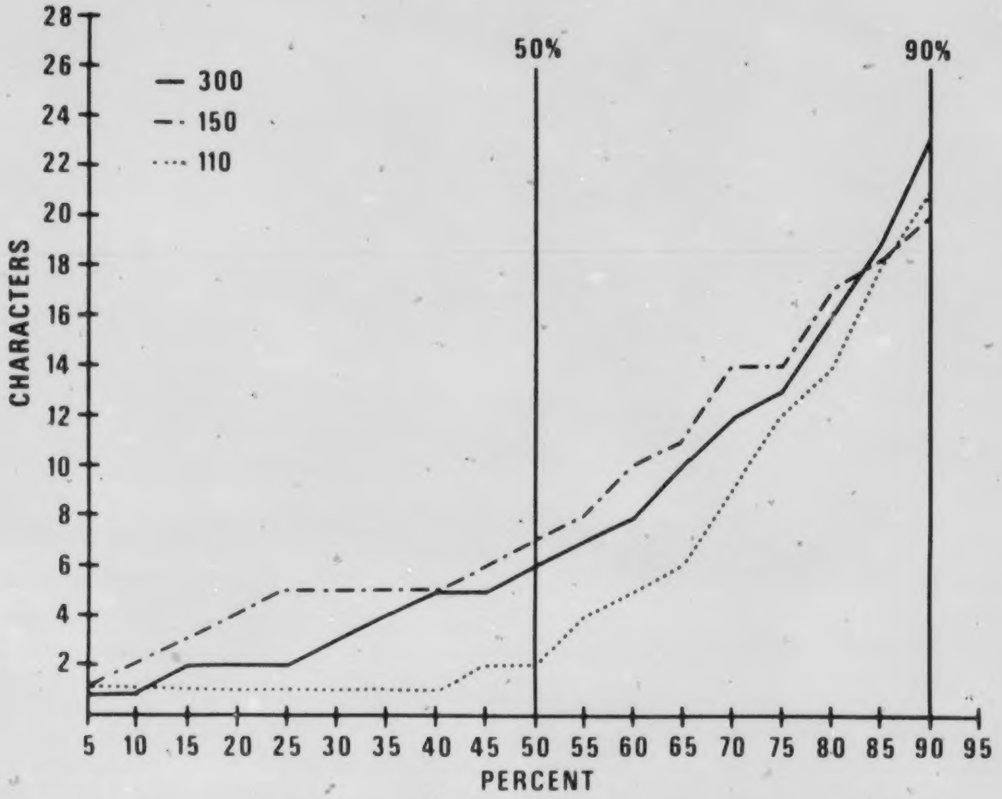


Figure 5-3. Cumulative Frequency Distribution of Stimulus Character Count

However, stimulus character count evidenced particularly high serial correlation (see Figure 4-1-c), so it is not surprising that the distribution of conversation medians do not show the same phenomenon. Thus, we must conclude that the trend shown for all observations is the result of a relatively few conversations with consistently higher stimulus character counts at the 300 bps data rate. It might be the case that users with applications requiring longer stimulus character counts tended to have longer sessions on higher speed terminals than on lower speed terminals, but this was not explicitly tested.

It was subsequently discovered (see section 5.3.4) that a number of users were inputting paper tape on lower speed terminals that were so equipped. This caused a dramatic change in the results for stimulus transmission rate that necessitated elimination of those message groups representing paper tape input. To determine if eliminating these message groups had any effect on stimulus character count, the data were trimmed at an upper limit of first 40 characters and then 80 characters. This was based on the presumption that paper tape input could be identified by the length of the input stream. In the first case (values for stimulus character count longer than 40 ignored) the number of observations for 110 bps data was reduced from 2638 to 2577 and the 90% value was reduced from 21 to 20. (The median was unchanged.) When the data were trimmed at 80, the number of observations was reduced to 2617 with similar effect on the median and 90% value. Since the trimming was shown to have negligible effect on the distribution of all observations, it was not performed for the distributions of conversation medians.

5.2.4 Stimulus Transmission Rate -

The stimulus transmission rate (Rs) is the rate in characters per second at which data are entered by the operator, measured after the first stimulus character has been entered. Thus, the stimulus transmission rate is derived by dividing the stimulus character count by the stimulus transmit time. (The response-stimulus delay time does not figure into the computations).

Measurements of stimulus transmission rate for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-4. The key percentile values for each class are tabulated in Table 5-5, both for the distribution of all observations and for the distribution of conversation medians. Note that the data presented in this table are expressed in characters per second, in order that the extra bit transmitted for each character at the 110 bps rate not distort comparisons of user data entry rates. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level without batching of the data, and they are significant at the 5% level when the data are batched (see section 5.5), though testing this data may not be relevent (see discussion below).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	5.2	10.0	2.3	9.9
150 BPS	3.0	6.0	3.0	3.1
300 BPS	3.4	30.0	3.0	5.7

Table 5-5. Key Percentiles for Stimulus Transmission Rate (Char/Sec)

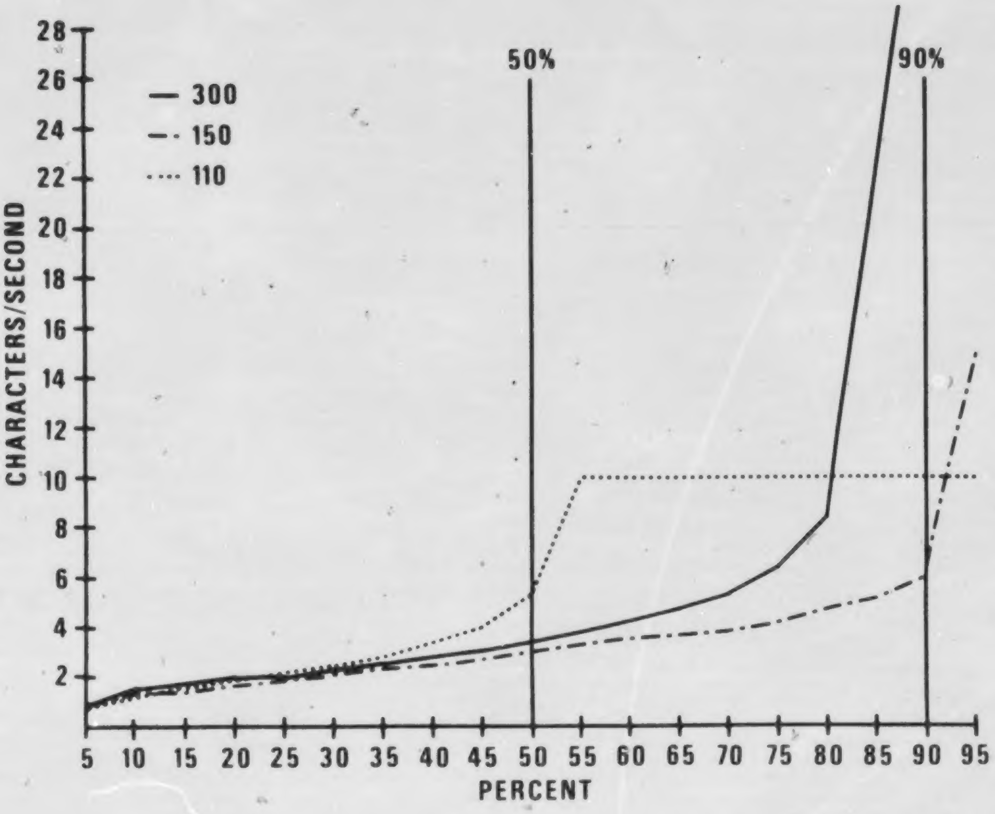


Figure 5-4. Cumulative Frequency Distribution of Stimulus Transmission Rate

Upon examination, these results were quite startling, since they implied user typing rates on low speed terminals that are intuitively unachievable. Further investigations led to the finding that (1) many stimuli whose transmission rate was at the maximum line rate (10, 15, or 30 cps) were single character inputs, and (2) some users were inputting paper tape which also had transmission rates at or near the maximum of 10 cps. (Also, the relatively high degree of serial correlation, as noted in Figure 4-1-d, indicates that users entering extremely short stimuli or paper tape inputs tended to do so throughout the conversation). Thus, while the user data entry rates may have been distributed as shown in Figure 5-4 and tabulated in Table 5-5, the user typing rates were not.

The simplest way to correct for the effects of single character stimuli and paper tape input was to trim all data above a given rate and recompute the percentiles. The decision to trim on the basis of rates is based on the maximum achievable human typing rates for similar situations (see section 2.1.1). The alternative would have been to examine each message group individually to determine if it was paper tape input. (Single character stimuli can readily be trimmed automatically). Examination of the percentiles revealed a jump from 5.2 to 10.0 between the 50 and 55 percentiles. Thus, it seemed reasonable to trim the data at 5 cps; i.e., all data values at or above 5.0 cps were eliminated from the analysis, and the percentiles were recompiled. This procedure was followed for all three terminal rates, resulting in the new values shown in Table 5-6. (The columns for "conversation medians" were computed from the distributions formed by trimming extreme values in each conversation prior to taking the median of the conversation).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	2.1	4.0	2.1	3.1
150 BPS	2.6	4.1	2.4	2.8
300 BPS	2.5	4.2	2.4	3.5

Table 5-6. Key Percentiles for Stimulus Transmission Rate After Elimination of Paper Tape Input and Single Character Stimuli (Char/Sec)

In the above table, the number of observations was reduced from 2638 to 1303 for 110 bps, from 361 to 295 for 150 bps, and from 19705 to 13211 for 300 bps. The number of observations for the distributions of conversation medians was naturally unchanged. Statistical tests of the distributions of medians of trimmed conversations showed the differences to be significant at the 1% level (see Table 5-22).

These results are more in keeping with previous findings presented in section 2.2.1*, and show a moderate increase in user typing rate with increased terminal speed. There are two possible explanations for this increase:

1. There is a general increase in the pace of operator interaction with the computer associated with the use of higher speed terminals.

---

* Note, however, that if these data are compared to the data reported in section 2.2.1, it must be kept in mind that these data are "burst" rates (measured from the start of the first stimulus character) while the data in section 2.2.1 are mean rates (including response-stimulus delay). See section 5.4.2 for mean rates that can be compared directly.

64

2. The keyboards provided with low speed teminals (110 bps - generally teletypwriters) is a limiting factor; providing typewriter-like keyboards results in higher user data entry rates.

There is no way with the present data to attribute the increase in stimulus character count to either of these explanations, although an increase in pace has already been noted in the results for response-stimulus delay time (section 5.2.1).

5.3 Effects Of Terminal Speed On System Performance Parameters

The system performance parameters are those associated with the acknowledgement and response portion of the message group. These parameters include the stimulus-acknowledgement delay time, the acknowledgement character count and transmission rate, the acknowledgement-response delay time, the response character count and transmission rate, and the stimulus-response delay time. (As previously noted, the acknowledgement and the response portion of the message group are not necessarily both present in every message group). The object here is to determine the effect on the computer system (as reflected by these parameters), if any, of varying the terminal speed.

5.3.1 Stimulus-Acknowledgement Delay Time -

The stimulus-acknowledgement delay time (D^a) is the interval between the transmission of the last charater of the stimulus and the receipt of the first character of the acknowledgement (if any). Since both the acknowledgement and the response are generated by the computer system, the distinction between these two components of the message group is determined by application of the SAR model, as described in previous chapters. This parameter is not defined for a particular message group if there is no acknowledgement.

Measurements of stimulus-acknowledgement delay time for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-5. The key percentile values for each class are tabulated in Table 5-7, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5). However, the differences between these distributions are all measured in units on the order of hundredths of a second, which are indistinguishable to users.

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	0.0	0.1	0.0	0.0
150 BPS	0.0	0.1	0.0	0.0
300 BPS	0.0	0.0	0.0	0.0

Table 5-7. Key Percentiles for Stimulus-Acknowledgement Delay Time (Seconds)

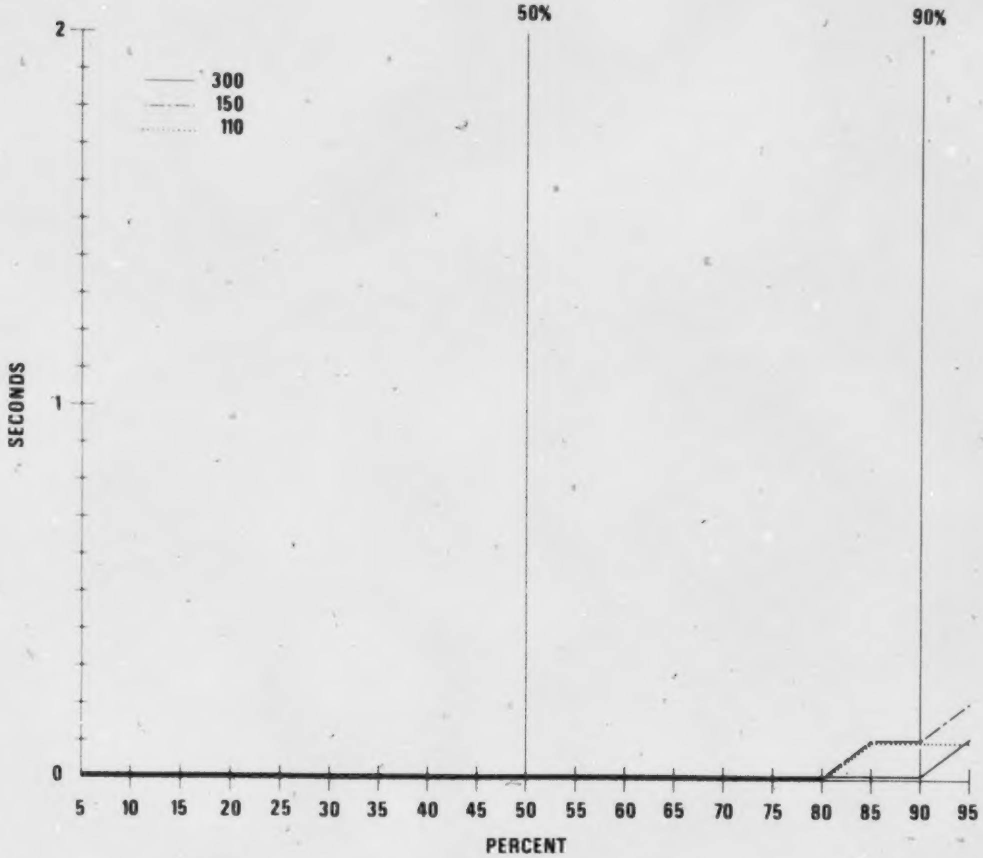


Figure 5-5. Cumulative Frequency Distribution of Stimulus -Acknowledgement Delay Time

Evidently, the computer system provided quite rapid acknowledgements to all interactive users, regardless of terminal speed. The differences between the distributions of stimulus-acknowledgement delay time, while statistically significant, have no practical significance.

5.3.2 Acknowledgement Transmit Time -

The acknowledgement transmit time (T^a) is the time from the first to the last character of the acknowledgement (if any). This parameter reflects both the system transmission rate for the acknowledgement and the number of characters in the acknowledgement. This time by itself is one component of the time required by the computer system for transmission on the channel to the user's terminal (the other component is the response transmit time). The acknowledgement transmission rate is obtained by dividing the acknowledgement character count (parameter 7) by the acknowledgement transmit time. This parameter is not defined for a particular message group if there is no acknowledgement.

Measurements of acknowledgement transmit time for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-6. The key percentile values for each class are tabulated in Table 5-8, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	0.2	0.9	0.2	0.9
150 BPS	0.7	1.7	0.3	0.7
300 BPS	0.2	0.7	0.3	0.3

Table 5-8. Key Percentiles for Acknowledgement Transmit Time (Seconds)

As for stimulus transmit time, acknowledgement transmit time reflects two opposite factors: acknowledgement character count and acknowledgement transmission rate. For the acknowledgement portion of the message group (see section 5.3.4), the transmission rate was essentially always at the maximum line rate, thus increasing directly with terminal speed. The acknowledgement character count (section 5.3.3) increased about as much at the median, but not at the 90-percentile, thus explaining the results for acknowledgement transmit time. In any event, the differences at the median are not likely to be perceptible to the user.

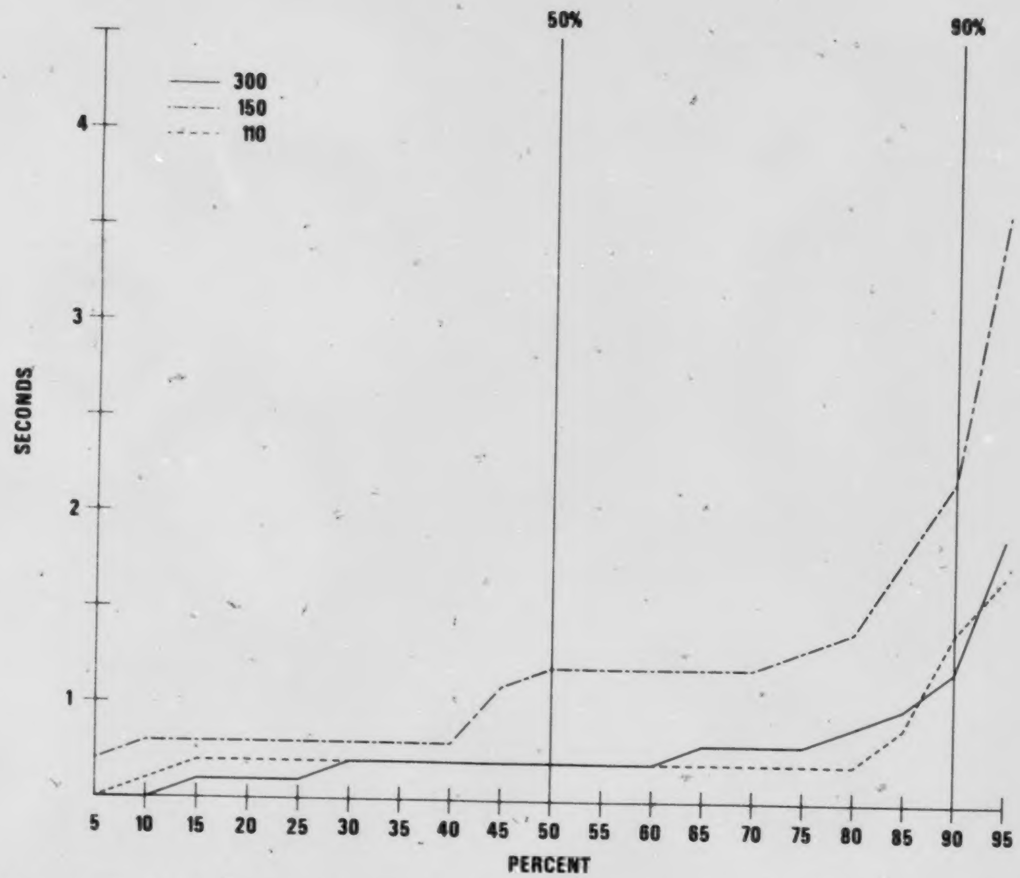


Figure 5-6. Cumulative Frequency Distribution of Acknowledgement Transmit Time

5.3.3 Acknowledgement Character Count -

The acknowledgement character count (C^a) is the number of characters in the acknowledgement (if any) from the computer system to the user terminal. This parameter is not defined for a particular message group if there is no acknowledgement.

Measurements of acknowledgement character count for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-7. The key percentile values for each class are tabulated in Table 5-9, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

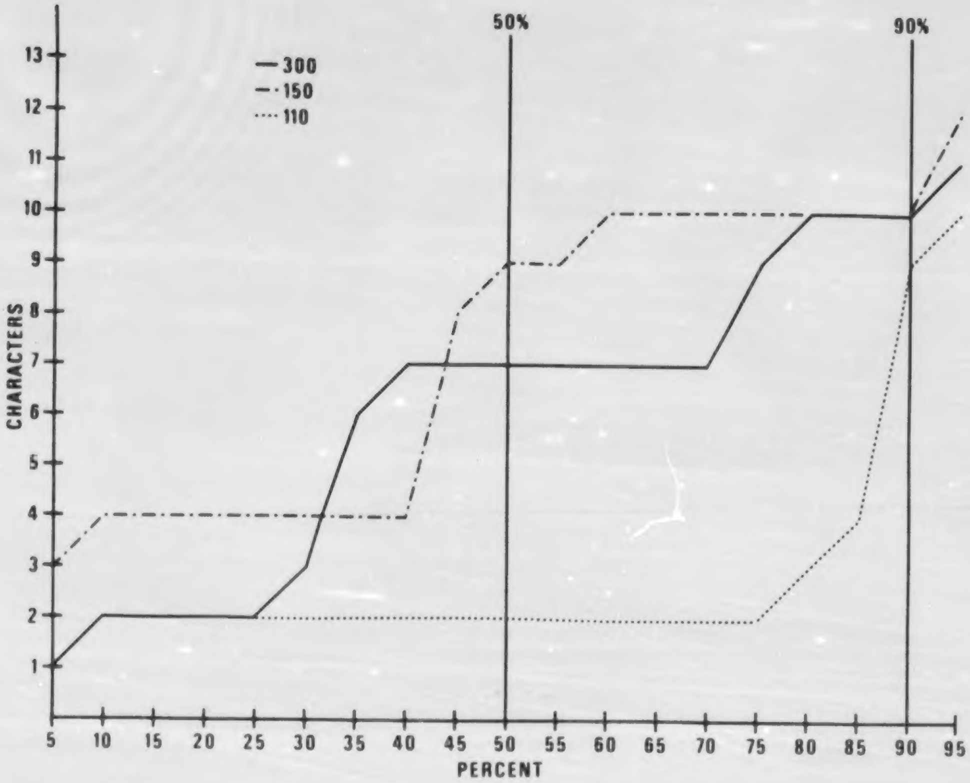


Figure 5-7. Cumulative Frequency Distribution of Acknowledgement Character Count

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	2	9	2	9
150 BPS	9	10	4	10
300 BPS	7	10	7	10

Table 5-9. Key Percentiles for Acknowledgement Character Count

The differences in acknowledgement character count appear to be due to the additional padding characters required for some (but not all) terminals operating at rates higher than 110 bps.

5.3.4 Acknowledgement Transmission Rate -

The acknowledgement transmission rate (R^a) is the rate in characters per second at which the acknowledgement (if any) is transmitted from the computer system to the user. This rate is computed by dividing the acknowledgement character count by the acknowledgement transmit time (the stimulus-acknowledgement delay time does not figure into the calculation). This parameter is not defined for a particular message group if there is no acknowledgement.

Measurements of acknowledgement transmission rate for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-8. The key percentile values for each class are tabulated in Table 5-10, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	10.0	10.0	10.0	10.0
150 BPS	15.0	15.0	15.0	15.0
300 BPS	30.0	30.0	30.0	30.0

Table 5-10. Key Percentiles for Acknowledgement Transmission Rate (Char/Sec)

Evidently, the short character strings in the acknowledgement were nearly always transmitted at burst rates. This explains the high degree of serial correlation for this parameter (evidenced in Figure 4-1-h). The relatively infrequent cases where the transmission was not at the maximum rate were probably due to output queue processing delays in the computer system. The impact of these infrequent delays on computer system line utilization is discussed in section 5.4.2.

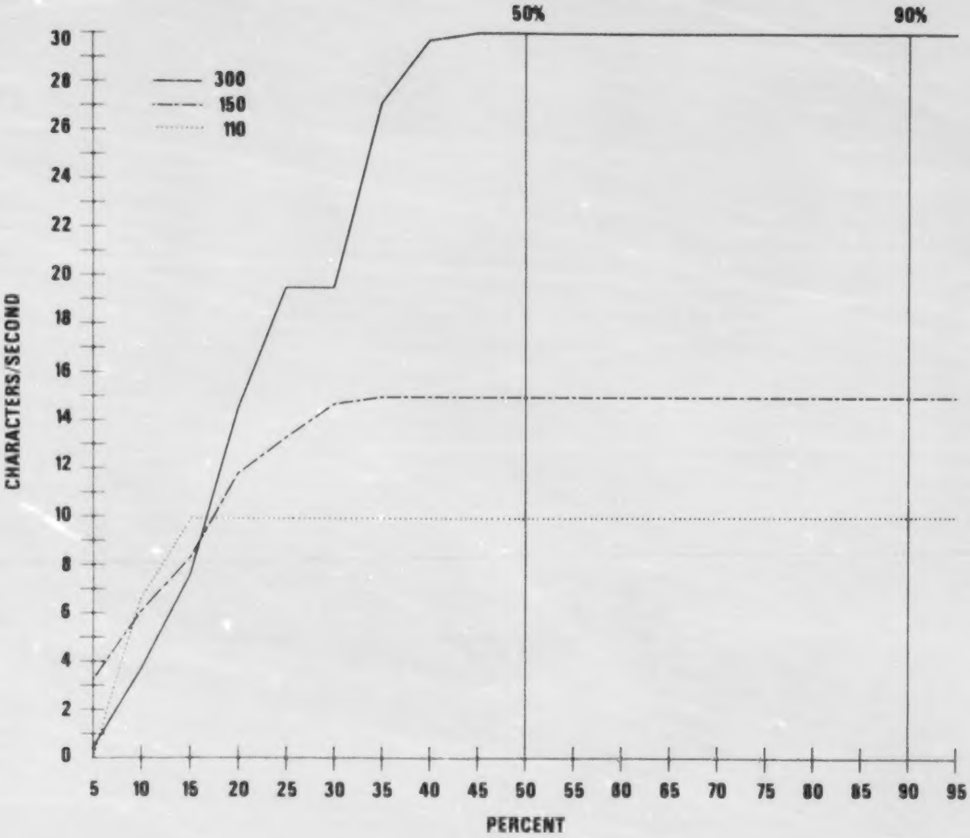


Figure 5-8. Cumulative Frequency Distribution of Acknowledgement Transmission Rate

5.3.5 Acknowledgement-Response Delay Time -

The acknowledgement-response delay time (D^r) is the idle time from the transmission of the last character of the acknowledgement (if any) to the first character of the response. Since both the acknowledgement and the response are generated by the computer system, the distinction between these two components of the message group is determined by application of the SAR model, as described in previous chapters. This parameter is not defined for a particular message group if there is no acknowledgement.

Measurements of acknowledgement-response delay time for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-9. The key percentile values for each class are tabulated in Table 5-11, both for the distribution of all observations and for the distribution of conversation medians.

71

The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	0.0	1.2	0.0	0.1
150 BPS	0.0	1.4	0.0	0.2
300 BPS	0.2	1.3	0.2	0.5

Table 5-11. Key Percentiles for Acknowledgement-Response Delay Time (Seconds)

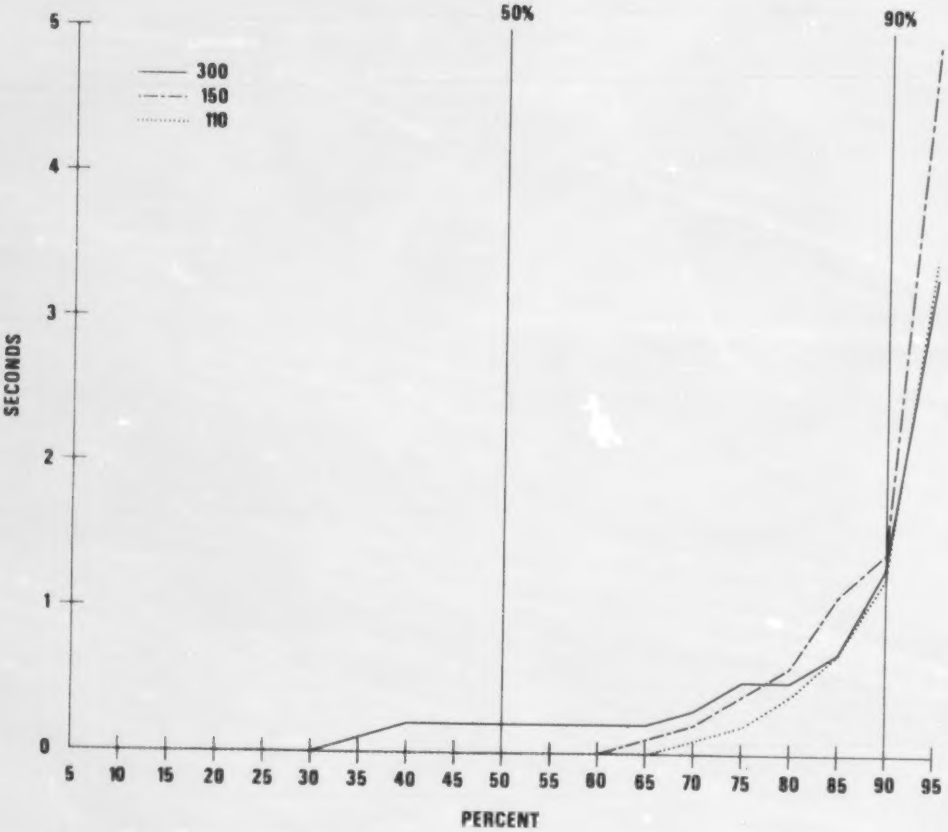


Figure 5-9. Cumulative Frequency Distribution of Acknowledgement-Response Delay Time

Clearly, the computer system in this study was providing quite rapid response to the interactive users (in addition to the nearly immediate acknowledgements evident in section 5.3.1). The data indicate slightly more rapid response to the users of 110 bps terminals than to the users of 300 bps terminals; however, the 0.2 seconds difference at the median is not likely to be perceptible to users, even though the difference between the distributions of delay time is statistically significant. This slight extra delay may reflect increased processing time required to set up the longer responses (noted in section 5.3.7) or more extensive application processing called for by the users' requests.

5.3.6 Response Transmit Time -

The response transmit time (T^r) is the time from the first to the last character of the response. This parameter reflects both the system transmission rate for the response and the number of characters in the response. By itself, this parameter is one component of the transmission time required by the computer system on the channel from the computer system to the user terminal. (The other component is the acknowledgement transmit time.) The response character count (parameter 11) is divided by the response transmit time to obtain the response transmission rate.

Measurements of response transmit time for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-10. The key percentile values for each class are tabulated in Table 5-12, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	0.2	9.3	2.5	3.9
150 BPS	2.2	7.8	2.1	2.7
300 BPS	1.1	6.5	1.1	2.1

Table 5-12. Key Percentiles for Response Transmit Time (Seconds)

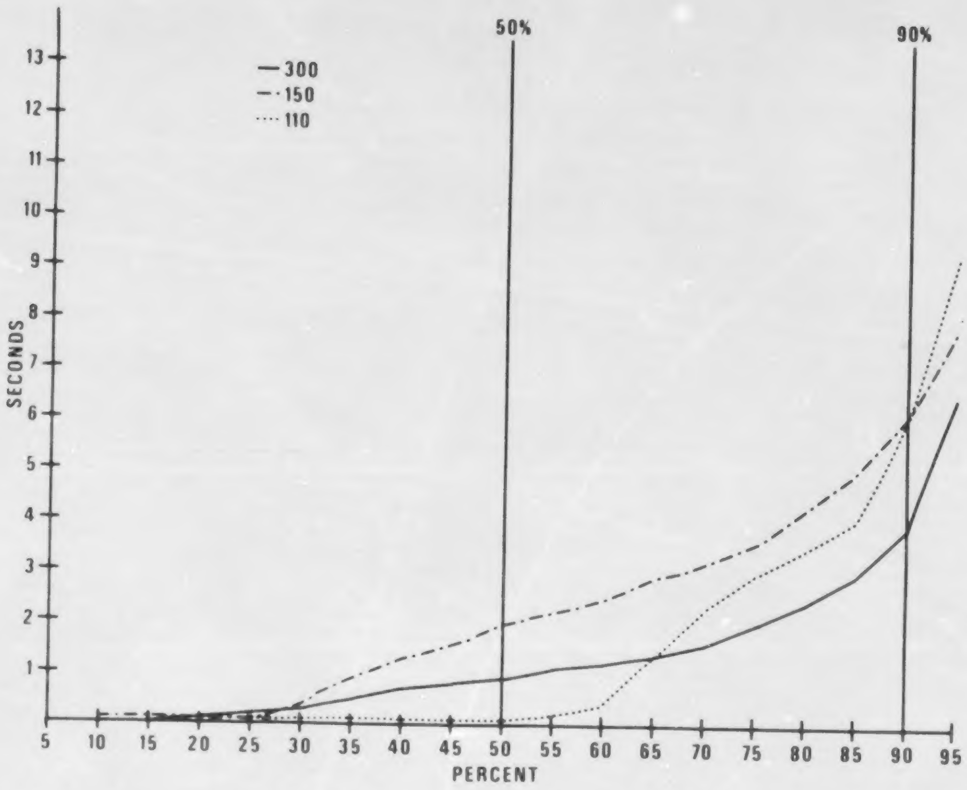


Figure 5-10. Cumulative Frequency Distribution of Response Transmit Time

As for stimulus transmission rate and acknowledgement transmission rate, response transmission rate reflects the net effect of two separate parameters: response character count and response transmission rate. Both response character count (section 5.3.7) and response transmission rate (section 5.3.8) clearly increase with increased terminal speed. However, when the effect of serial correlation is eliminated, it is evident that the effect of increased response transmission rate dominates, since response transmit time decreases despite the increase in response character count. The low 50-percentile value for all observations of response transmit time for the 110 bps class probably reflects the effect of a number of conversations with consistently brief responses.

4

5.3.7 Response Character Count -

The response character count (C^r) is the number of characters in the response transmitted from the computer system to the user terminal. Since both the acknowledgement and the response are generated by the computer system, the distinction between these two components of the message group is determined by application of the SAR model, as described in previous chapters.

Measurements of response character count for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-11. The key percentile values for each class are tabulated in Table 5-13, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level without batching of the data, and they are significant at the 5% level when the data are batched (see section 5.5).

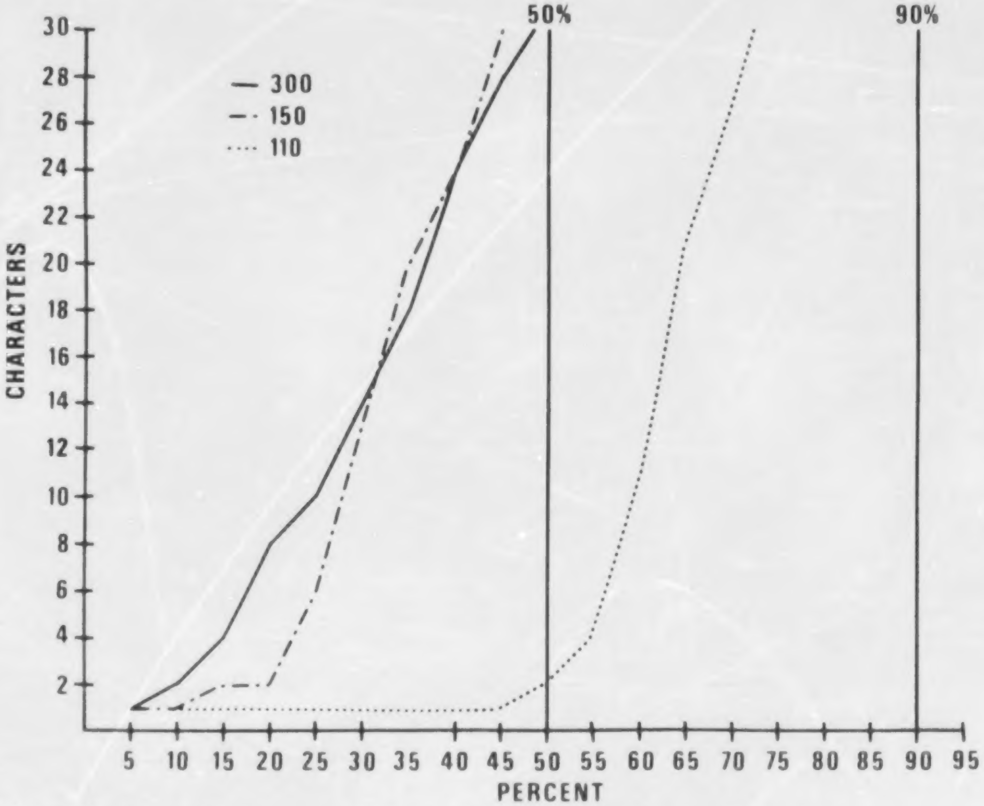


Figure 5-11. Cumulative Frequency Distribution of Response Character Count

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	2	91	25	39
150 BPS	33	117	32	41
300 BPS	31	159	30	60

Table 5-13. Key Percentiles for Response Character Count

The increase in response character count noted for increased terminal speed is the key result of the investigation into the effect of terminal speed on system performance parameters. This increase parallels the previously noted increase in stimulus character count as a function of terminal speed when all observations are considered. (No increase was found when the data were batched). Actually, of course, the increase is not attributable to the system itself, but to the nature of the service requests generated by the user. Apparently, at higher speeds some users are more verbose in their inputs and most request outputs in more verbose form as well. To the extent that longer responses contain more or more easily understood information, this increase can be interpreted as an increase in response quality, as well as quantity.

One effect of this increase in response character count with increased terminal speed is that the response transmission time does not decrease as much as would be expected with increased line speed. Disregard of this phenomenon could lead to under-estimation of required line capacity in terminal multiplexor or concentrator design.

5.3.8 Response Transmission Rate -

The response transmission rate (R^r) is the rate in characters per second at which the response is transmitted from the computer system to the user terminal, measured after the first response character has been transmitted. Thus, the response transmission rate is derived by dividing the response character count by the response transmit time. (The acknowledgement-response delay time, if any, does not figure into the computation).

Measurements of response transmission rate for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-12. The key percentile values for each class are tabulated in Table 5-14, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	10.0	10.0	9.9	10.0
150 BPS	15.0	15.0	15.0	15.0
300 BPS	30.0	30.0	30.0	30.0

Table 5-14. Key Percentiles for Response Transmission Rate (Char/Sec)

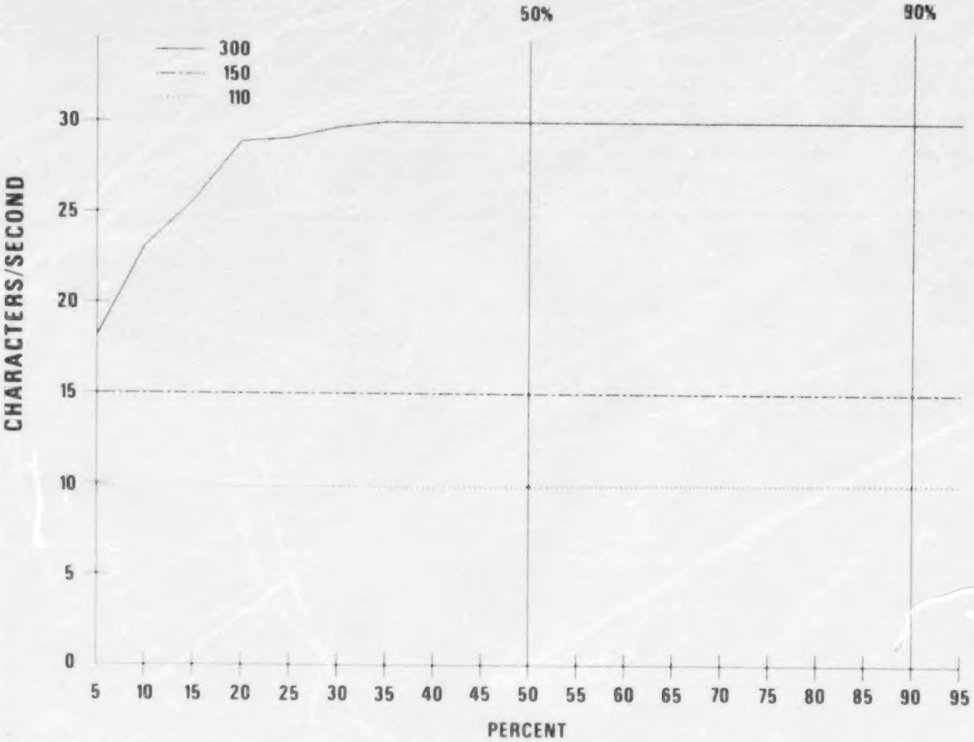


Figure 5-12. Cumulative Frequency Distribution of Response Transmission Rate

Even with the longer character strings in the response (as compared to the acknowledgement), the speed of data transmission of the response tended to be the maximum possible for all terminal speeds. This explains the high degree of serial correlation for this parameter (evidenced by Figure 4-1-1). Thus, the transmission of neither the acknowledgement nor the response tended to be very bursty. This can be attributed to the fact the interactive usage of the computer system under investigation was limited during the period of the study (coincidently) so that system responsiveness to those users who were permitted on was quite good.

5.3.9 Stimulus-Response Delay Time -

The stimulus-response delay time (D_{sr}) is the time from the last character of the stimulus to the first character of the response. If an acknowledgement is present in the message group, this time is the sum of the stimulus-acknowledgement delay time, the acknowledgement transmit time and the acknowledgement-response delay time (see equation 5, section 2.3.3.2.1). If no acknowledgement is present in the message group, this time is idle time. As explained in previous chapters, this parameter is a measure of the service time for the computer system to deliver information relevant to the query to the terminal operator.

Measurements of stimulus-response delay time for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-13. The key percentile values for each class are tabulated in Table 5-15, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

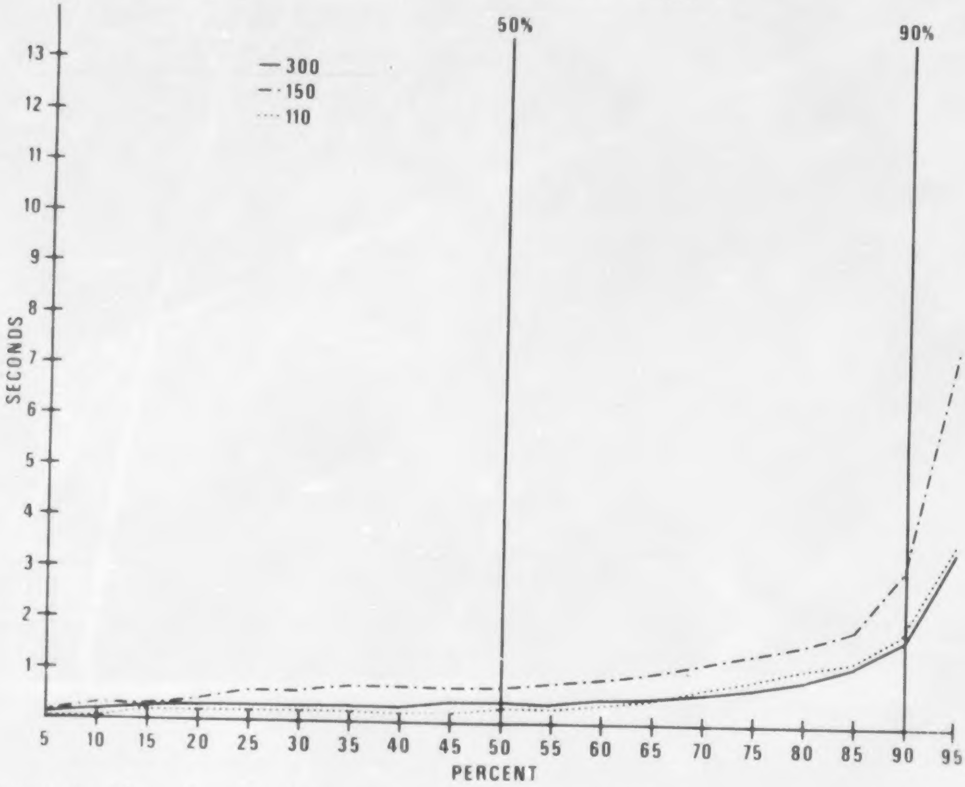


Figure 5-13. Cumulative Frequency Distribution of Stimulus-Response Delay Time

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	0.3	1.9	0.3	1.0
150 BPS	0.7	3.1	0.6	0.8
300 BPS	0.4	1.8	0.4	0.8

Table 5-15. Key Percentiles for Stimulus-Response Delay Time (Seconds)

Since the stimulus-acknowledgement delay time was so small for all cases, this parameter is effectively the sum of the acknowledgement transmit time and the acknowledgement-response delay time. The differences between the distributions, though statistically significant, are not likely to be perceptible to the user. As noted several times previously, the service provided to interactive users was remarkably good during the period of the study since the interactive load was kept quite limited during this time.

5.4 Overall Effects Of Varying Terminal Speed

The overall effects of varying the terminal speed on both the user and the interactive computer system are reflected in the stimulus inter-arrival time parameter of the SAR model and in line utilization statistics that are accumulated for the user (stimulus) and system (acknowledgement and response) portions of transmission traffic on the communications line.

5.4.1 Stimulus Inter-arrival Time -

The stimulus inter-arrival time (D^{SS}) is a parameter that reflects both user and system performance parameters. This time is measured from the time of entry of the first character of a stimulus to the time of entry of the first character of the successive stimulus. (This time is not defined for the first message group in a conversation). In a sense, it is the "duty cycle" for interactive use of the computer system, since it measures the time between successive service requests. The stimulus inter-arrival time may be computed by adding the stimulus transmit time, the stimulus-acknowledgement delay time (if any), the acknowledgement transmit time (if any), the acknowledgement-response delay time (if any, or else the stimulus-response delay time), the response transmit time, and the response-stimulus delay time (to the next successive stimulus) (see equation 6, section 2.3.3.2.1).

Measurements of stimulus inter-arrival time for all message groups, categorized by terminal speed, were distributed as shown in Figure 5-14. The key percentile values for each class are tabulated in Table 5-16, both for the distribution of all observations and for the distribution of conversation medians. The differences between the distributions for the 110 bps and the 300 bps classes are significant at the 1% level, both with and without batching of the data (see section 5.5).

	All Observations		Conversation Medians	
Rate	50%	90%	50%	90%
110 BPS	5.5	37.8	13.3	21.2
150 BPS	10.7	32.5	10.1	12.7
300 BPS	7.4	34.8	8.9	18.0

Table 5-16. Key Percentiles for Stimulus Inter-arrival Time (Seconds)

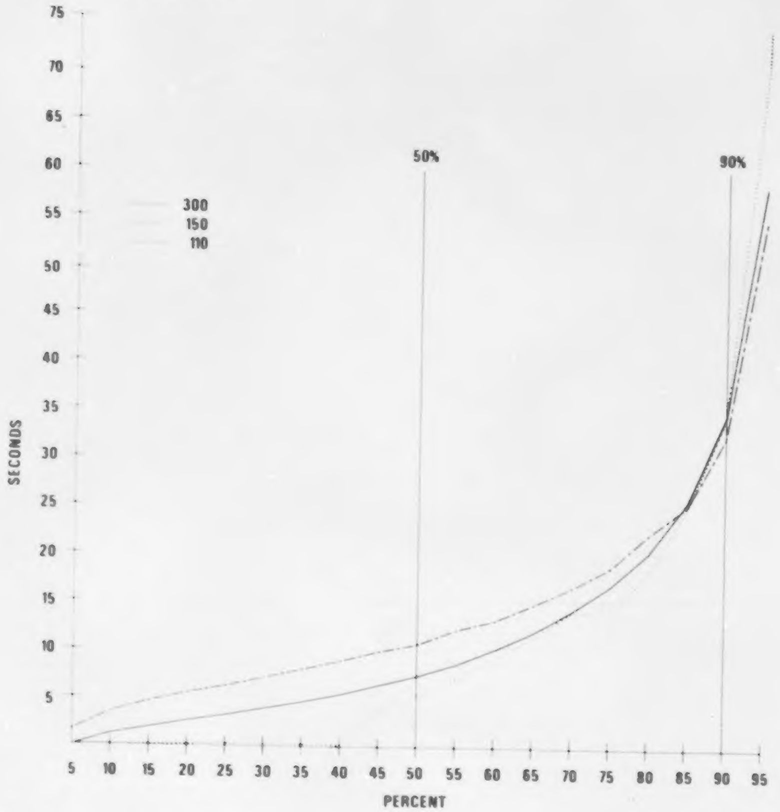


Figure 5-14. Cumulative Frequency Distribution of Stimulus Inter-Arrival Time

Stimulus inter-arrival time did exhibit some degree of serial correlation (see Figure 4-1-n). After removal of this serial correlation, there is evidence of an increase in the pace of interaction with increased terminal speed.

It may be of interest to examine the components of stimulus inter-arrival time to determine their relative impact on the total time. We consider the mean values here, since, for an equal number of observations for all components, the sum of the individual means should equal the mean of the sum (or stimulus inter-arrival time itself). The mean values of the components of stimulus inter-arrival time are tabulated in Table 5-17. However, since extreme values can unduely affect the mean, the data for each parameter have been trimmed to eliminate observations greater than 60 seconds. The empirical mean values of (trimmed) stimulus inter-arrival time presented in Table 5-17 do not equal the sum of the means of all the components of stimulus inter-arrival time due to the unequal effects of the trimming operation on the means of all the components.

It is evident from Table 5-14 that for the present investigation, the contribution of the computer's acknowledgement (D^a, T^a) and delay before response (D^r) are inconsequential. User think time (D^s) and stimulus entry time (T^s) are the major components, though the computer response (T^r) also occupies significant time.

	110 BPS	150 BPS	300 BPS
D^s	6.0	7.2	4.3
T^s	4.3	6.3	4.8
D^a	0.1	0.2	0.1
T^a	0.4	0.8	0.4
D^r	0.7	0.9	0.8
T^r	3.0	3.8	2.8
D^{ss} (Total)	9.2	13.6	10.9

Table 5-17. Mean Values of Components of Stimulus Inter-Arrival Time

5.4.2 Line Utilization -

The Data Analysis Package includes the capability to compute certain line utilization statistics for sets of data, including the total number of printing and non-printing charaters transmitted by the user and the system, and the percentage of transmission capacity utilized by the user and by the system, both including and excluding delay times prior to transmission. These statistics are tabulated for the 110 bps, 150 bps and 300 bps data in Tables 5-18, 5-19 and 5-20, respectively.

	Number	% (total)	% (subtotal)
Total Number of characters	210499	100.0	
User	43182	20.5	100.0
Printing	38902	18.5	90.1
Non-printing	4280	2.0	9.9
System	167317	79.5	100.0
Printing	150025	71.3	89.7
Non-printing	17292	8.2	10.3
Subtotal of printing characters	188927	89.8	100.0
User	38902	18.5	20.6
System	150025	71.3	79.4
Subtotal of non-printing chars	21572	10.2	100.0
User	4280	2.0	19.8
System	17292	8.2	80.2
			Rate Char/Sec
Burst Line Utilization			
stimulus		20.5%	2.1
acknowledgement		91.5%	9.2
response		99.5%	10.0
Mean Line Utilization			
stimulus		11.2%	1.1
system		86.7%	8.7

Table 5-18. Line Utilization of 110 BPS Conversations

Comparing the percentage of total characters sent by the user, a decrease from 20.5% at 110 bps to 11.5% at 300 bps may be noted. (The 150 bps data are anomolous due to the small number of observations). This reflects the more verbose responses called for by the users of higher speed terminals (since the ratio of system characters to total characters sent increases with the speed of the terminal).

Examining the ratio of printing to non-printing characters for the user and the system, we find that while it has not changed much for the user, there is a noticeable increase in the percentage of non-printing characters output by the computer. This may be indicative of less densly packed output to the user (e.g., more liberal use of blank spaces and blank lines to enhance readability) or simply the need to use more padding characters (to allow time for physical carriage returns) with the higher speed terminals. Probably both factors contribute.

	Number	% (total)	% (subtotal)
Total number of characters	30885	100.0	
User	7146	23.1	100.0
Printing	5880	19.0	82.3
Non-printing	1266	4.1	17.7
System	23739	76.9	100.0
Printing	16514	53.5	69.6
Non-printing	7225	23.4	30.4
Subtotal of printing characters	22394	72.5	100.0
User	5880	19.0	26.3
System	16514	53.5	73.7
Subtotal of non-printing chars	8491	27.5	100.0
User	1266	4.1	14.9
System	7225	23.4	85.1

		Rate Char/Sec
Burst Line Utilization		
stimulus	15.0%	2.3
acknowledgement	69.5%	10.4
response	85.4%	12.8
Mean Line Utilization		
stimulus	7.6%	1.1
system	75.0%	11.3

Table 5-19. Line Utilization of 150 BPS Conversations

Burst line utilization statistics reflect the effective transmission rate of the user and the system after each has begun transmitting (thus, idle time latencies prior to transmission are ignored). We find user burst rates of 2.1 cps, 2.3 cps and 2.3 cps for the 110 bps, 150 bps and 300 bps data, respectively. These data may be compared with the various input data rates cited in section 2.1.1 and with the stimulus transmission rates given in section 5.2.4 (though cumulative percentile values are given there, as opposed to mean rates here). There appears to be a slight improvement in user transmission rate for the 150 and 300 bps terminals, as compared with the 110 bps terminals. This may be attributable to the type of keyboard used, since the 110 bps terminals are all teletypewriters with a significantly different keyboard "feel" from the typewriter-like keyboards on the 150 and 300 bps terminals.

	Number	% (total)	% (subtotal)
Total number of characters	2190232	100.0	
User	252009	11.5	100.0
Printing	224275	10.2	89.0
Non-printing	27734	1.3	11.0
System	1938223	88.5	100.0
Printing	1493903	68.2	77.1
Non-printing	444320	20.3	22.9
Subtotal of printing characters	1718178	78.4	100.0
User	224275	10.2	13.1
System	1493903	68.2	86.9
Subtotal of non-printing chars	472054	21.6	100.0
User	27734	1.3	5.9
System	444320	20.3	94.1

		Rate Char/Sec
Burst Line Utilization		
stimulus	7.7%	2.3
acknowledgement	53.6%	16.1
response	61.5%	18.5
Mean Line Utilization		
stimulus	4.2%	1.3
system	52.9%	15.9

Table 5-20. Line Utilization of 300 BPS Conversations

Looking at burst statistics for system output, we find a increasing actual rate but a decreasing utilization of line capacity as we progress from the lower to higher speed terminals. (Again, it may be interesting to compare these rates with the acknowledgement transmission rates given in section 5.3.4 and the response transmission rates given in section 5.3.8, keeping in mind that cumulative percentile values are presented there in contrast to mean rates here). Apparently, the servicing of output queues is performed so sporadically that the computer is unable to output continuously at the maximum line rate.

Mean utilization statistics do include the latencies (think time or processing time) prior to actual transmission. For the user, we find the rates to be 1.1 cps, 1.1 cps and 1.3 cps for the 110 bps, 150 bps and 300 bps terminals, respectively. These are the appropriate rates to compare with the data presented in section 2.2.1 (particularly the Bell Laboratories' data in Table 2-1), since that data also included user delay time. The input rates observed in the present study are higher than in the Bell Laboratories' study at the lower speed of 110 bps, but are slightly lower at the 150 bps rate and are slightly higher at the 300 bps rate (though no direct comparison can be made at the latter rate). The results are not conclusive, but the increase at the 110 bps rate may be attributable to the relatively light loading for interactive users on the NBS system during the study period. This would be in keeping with the Bell Laboratories finding that user typing rates were higher on a moderately loaded system than on a heavily loaded system.

Mean utilization data for the system show a similar trend as for system burst utilization: increasing absolute transmission rates with higher speed terminals, but decreasing line utilization. The mean transmission rates were 8.7 cps, 11.3 cps and 15.9 cps for the 110, 150 and 300 bps cases, respectively. Line sharing techniques are most indicated for higher speed terminals (since, for example, the 300 bps case had only 52.9% utilization of the line by the system and only 4.2% by the user).

5.5 Statistical Significance

In this section, we discuss in somewhat greater detail the application of the statistical procedures described in Chapter 4 to the parameter distributions presented in this chapter. As discussed in Section 4.2, commonly used tests of significance are inappropriate for use with these data, since the assumption of an underlying normal distribution is invalid. For this reason, computer programs were written by the author (using subroutines in the IBM Scientific Subroutine Package) to apply two non-parametric tests to the data. These tests are the Mann-Whitney U-test (discussed in Section 4.2.1) and the Kolmogorov-Smirnov Test (discussed in Section 4.2.2). Table 5-21 presents the results of applying these two tests to determine the significance of the differences between the distributions of parameter values of the 300 bps data and the 110 bps data, for each SAR model parameter. The significance levels associated with the various values of Z in this table are not presented, since they are all better than $\alpha < .001$. This is probably due to the large sample sizes involved -- with such large samples, even small differences between the sample distributions indicate a consistent difference in the underlying distributions.

The sample size for the 150 bps data was not considered adequate and consequently these data were not tested against either the 110 bps or the 300 bps data.

Parameter	Kolmogorov-Smirnov	Mann-Whitney	
	Z	U	Z
D^s	7.13	1.00×10^7	-14.45
T^s	13.82	2.24×10^7	-11.27
C^s	13.95	1.98×10^7	-19.70
R^s	13.75	2.33×10^7	- 8.53
D^a	7.22	1.09×10^7	- 6.63
T^a	19.57	1.11×10^7	- 9.46
C^a	19.48	0.67×10^7	-31.40
R^a	31.75	0.47×10^7	-49.20
D^r	13.65	0.45×10^7	-18.22
T^r	16.18	1.20×10^7	- 9.46
C^r	18.79	1.31×10^7	-33.37
R^r	46.20	0.10×10^7	-85.70
D^{sr}	13.10	1.12×10^7	-10.77
D^{ss}	16.41	2.07×10^7	-15.37

Table 5-21. Results of Non-Parametric Tests Between 110 BPS and 300 BPS for All Parameters Using All Data (Number of Observations Given as Given in Table 5-1)

The question of data independence was discussed in section 4.1.2. In order to investigate the possibility of serial correlation between successive observations in the same convesation, the autocorrelation coefficients for all parameters were computed for each conversation, and the distribution of autocorrelation coefficients (over all conversations) was displayed for each parameter. Those results suggested the existence of some degree of serial correlation in some parameters, particularly those associated with the stimulus portion of the message group.

As previously described, batching of observations is a method for eliminating serial correlation. The median of each conversation was chosen as an appropriate statistic for this purpose. After collecting the medians of all parameters for each conversation into a file, the Kolmogorov-Smirnov test and the Mann-Whitney U-test were then run on the distributions of the medians of each parameter for the 110 bps and 300 bps classes. The results of these tests are summarized in Table 5-22.

It is evident from this table that the differences between the distributions of most parameters is still statistically significant at very high levels ($\alpha < .01$). However, two parameters of high interest, stimulus character count and response character count, fail to be significantly enough different for the null hypothesis to be rejected at the 1% level (though response character count did not seem to exhibit a high degree of serial correlation, as shown in Figure 4-1-k).

Parameter	Kolmogorov-Smirnov		Mann-Whitney		
	Z	P(Z)	U	Z	P(Z)
D^s	1.65	.009	2699	-3.05	.001
T^s	1.69	.007	3198	-1.89	.030
C^s	1.01	.257	3673	-0.79	.216
R^s	1.46	.028	3089	-2.14	.016
R^{s}*	1.47	.026	2900	-2.58	.005
D^a	2.11	.000	3008	-2.05	.020
T^a	2.11	.000	2366	-3.82	.000
C^a	3.00	.000	2102	-4.71	.000
R^a	4.84	.000	807	-7.91	.000
D^r	1.65	.009	2699	-3.05	.001
T^r	2.90	.000	2547	-3.41	.000
C^r	0.94	.340	3137	-2.03	.021
R^r	5.39	.000	0	-10.2	.000
D^{sr}	2.11	.000	3044	-2.24	.012
D^{ss}	2.35	.000	2607	-3.26	.001

Table 5-22. Results of Non-Parametric Tests Between 110 BPS and 300 BPS for All Parameters Using Medians of Each Message Group (Number of Observations as Given in Table 2-3)

---

* Trimmed as per Table 5-6.

6.0 EFFECTS OF DIFFERENCES IN APPLICATION

In this chapter we discuss the results of analyzing the data according to application or type of use of each message group. This is an independent grouping scheme from that employed in Chapter 4 and in principle, a two-way analysis could be performed to determine the effects of each class or "treatment" simultaneously. However, in order to simplify the analysis (since it is not the main thrust of this dissertation), only the data collected at 300 bps were grouped and analyzed by application. Since the volume of data collected for terminals operating at 300 bps was considerably greater than for data collected for terminals operating at other speeds, such a limited analysis is still adequate to illustrate how this measurement and analysis approach may be applied to compare in a quantitative way differences in user and system performance when the computer system is used for different tasks.

6.1 Subsets Defined

The nature of subsets and the subset assignment process were discussed in Sections 3.1.2.3 and 3.3.2, respectively. The software that scans the text of the message groups (actually the stimulus portion only) to assign them to subsets can identify the following subsets for use of the Univac 1108 computer:

ADD - includes a file in the command stream
ALG - invokes the ALGOL compiler
ASG - assigns peripheral devices
ASM - invokes the assembler
BASIC - invokes the BASIC interpreter
BRKPT - breakpoint (writes a core memory image for restart)
CFOR - invokes the conversational FORTRAN interpreter
COB - invokes the COBOL compiler
FREE - deallocates a file or peripheral device
FUR - various File Utility Routines commands
DATA - inserts a data file into the command stream
ED - invokes the conversational editor
ELT - creates a file
FOR - invokes the FORTRAN compiler
FIN - terminates a run
LIST - causes a file to be printed
MAP - invokes the link editor
PMD - causes a post-mortum dump (in event of a job "crash")
RALPH - invokes the RALPH (dialect of FORTRAN) interpreter
RFOR - invokes the reentrant FORTRAN interpreter
RUN - initiates a job or conversational session
START - executes a user program stored in a file
STAT - causes a system status message to be printed
SYM - directs files to peripheral devices
USE - assigns additional names to files
XBAS - invokes the XBAS (dialect of BASIC) interpreter
XQT - executes a program

6.2 Subset Assignment

The subset assignment program accumulates statistics on the number of message groups and the number of characters in those message groups (including stimulus, acknowledgement and response portions) for each subset. These statistics include all the conversations for which assignments are made in a single processing run. By assigning subsets for all conversations at the same line speed in three separate runs, summary statistics of subset usage by speed group may readily be obtained. Tables 6-1, 6-2 and 6-3 present these summary statistics of subset assignment for conversations at 110 bps, 150 bps and 300 bps, respectively. The summary of subset assignments for 110 bps and 150 bps data are presented here for information and comparison; however, these data are not analyzed further.

It is interesting to note that the percentage of message groups on tasks requiring sustained, "intimate" interaction with the computer system, such as editing, increase with the increase in terminal speed, while the percentage of message groups characteristic of setting up remote batch-type jobs, such as program execution, decrease with the increase in terminal speed. This may be interpreted as further evidence that user prefer higher speed terminals for highly interactive tasks.

6.3 Tabulation Of Summaries

Table 6-4 tabulates the 50-percentile (median) and 90-percentile values for each of the fourteen SAR model parameters for all 300 BPS conversations, grouped by application subset. The number of observations (message groups) for each parameter for each subset are included in the table.

6.4 Sample Analyses

The data presented in Table 6-4 are rich in possibilities for analysis; so rich, in fact, that no single chapter would suffice to discuss even a good percentage of the possibilities. Since the concern here is more with illustrating the methodology and the possible range of of its application, only a few of the more interesting possible analyses are discussed.

SUBSET	CHARACTERS		MESSAGE GROUPS	
ADD	866	1.64%	101	3.83%
ASG	776	1.47%	44	1.67%
BRKPT	193	0.37%	9	0.34%
FUR	1570	2.97%	76	2.88%
DATA	3674	6.95%	694	26.31%
ED	5888	11.15%	363	13.76%
ELT	616	1.17%	18	0.68%
FIN	1017	1.93%	29	1.10%
FOR	2966	5.61%	87	3.30%
FREE	93	0.18%	6	0.23%
MAP	995	1.88%	63	2.39%
RALPH	33	0.06%	1	0.04%
RUN	1447	2.74%	37	1.40%
START	91	0.17%	4	0.15%
STAT	279	0.53%	14	0.53%
SYM	72	0.14%	4	0.15%
USE	72	0.14%	3	0.11%
XBAS	109	0.21%	6	0.23%
XQT	28062	53.12%	767	29.08%
NO SUBSET	4011	7.59%	312	11.83%
	-----		----	
Total	52830		2638	

Table 6-1. Subset Assignment Summary for 110 BPS Conversations

SUBSET	CHARACTERS		MESSAGE GROUPS	
ADD	182	1.49%	6	1.66%
ASG	340	2.78%	12	3.32%
BRKPT	438	3.58%	14	3.88%
FUR	834	6.81%	39	10.80%
ED	2114	17.27%	86	23.82%
ELT	525	4.29%	20	5.54%
FIN	296	2.42%	7	1.94%
FOR	116	0.95%	4	1.11%
FREE	225	1.84%	8	2.22%
MAP	372	3.04%	13	3.60%
PMD	24	0.20%	1	0.28%
RUN	354	2.89%	8	2.22%
SYM	243	1.99%	8	2.22%
XQT	5740	46.90%	105	29.09%
NO SUBSET	437	3.57%	30	8.31%
	-----		---	
Total	12240		361	

Table 6-2. Subset Assignment Summary for 150 BPS Conversations

SUBSET	CHARACTERS		MESSAGE GROUPS	
ADD	9261	1.92%	497	2.52%
ASG	10663	2.21%	396	2.01%
ASM	141	0.03%	7	0.04%
BASIC	914	0.19%	28	0.14%
BRKPT	2940	0.61%	131	0.66%
COB	692	0.14%	29	0.15%
DATA	4756	0.99%	247	1.25%
ED	122808	25.44%	5846	29.67%
ELT	6894	1.43%	186	0.94%
FIN	10596	2.20%	227	1.15%
FOR	2712	0.56%	93	0.47%
FREE	4617	0.96%	215	1.09%
FUR	26578	5.51%	1036	5.26%
LIST	43	0.01%	2	0.01%
MAP	3613	0.75%	163	0.83%
PMD	703	0.15%	93	0.47%
RALPH	520	0.11%	20	0.10%
RFOR	213	0.04%	10	0.05%
RUN	10007	2.07%	238	1.21%
START	3091	0.64%	98	0.50%
STAT	5697	1.18%	294	1.49%
SYM	680	0.14%	37	0.19%
USE	3936	0.82%	131	0.66%
XBAS	12008	2.49%	670	3.40%
XQT	216757	44.90%	7929	40.24%
NO SUBSET	21870	4.53%	1082	5.49%
Total	482710		19705	

Table 6-3. Subset Assignment Summary for 300 BPS Conversations

SUBSET	D^s			T^s			C^s		
	#Obs	50%	90%	#Obs	50%	90%	#Obs	50%	90%
ADD	456	0.2	2.4	497	0.4	7.4	497	2	18
ASG	393	0	3.9	396	6.3	15.8	396	15	22
ASM	7	0	0	7	7.4	9.3	7	14	19
BASIC	29	3.0	27.8	28	21.2	47.7	28	20	25
BRKPT	131	0.2	2.9	131	5.3	9.4	131	14	20
COB	17	0.3	2.9	17	4.3	12.8	17	14	16
DATA	160	0	1.8	247	0	13.0	247	1	16
ED	5707	1.6	12.7	5846	2.2	13.6	5846	6	21
ELT	186	1.1	6.1	186	0.7	10.8	186	20	32
FIN	227	0	0	227	1.5	3.2	227	5	5
FOR	90	0	1.5	93	6.2	14.9	93	16	27
FREE	194	0	1.4	215	3.8	7.8	215	11	17
FUR	976	1.0	13.4	1036	5.8	19.5	1036	14	27
LIST	2	0	0	2	8.8	8.8	2	16	16
MAP	161	0.7	2.7	163	5.2	11.9	163	12	21
PMD	12	1.6	6.4	93	0	8.0	93	1	13
RALPH	18	0	0.7	20	7.9	16.7	20	19	34
RFOR	10	0	0	10	6.9	9.1	10	18	18
RUN	236	0	3.5	238	11.0	23.2	238	31	41
START	98	0	0.8	98	8.9	25.0	98	19	28
STAT	252	0	63.6	294	2.3	5.1	294	6	14
SYM	37	0	1.2	37	3.6	10.2	37	10	21
USE	131	0	0.8	131	7.0	13.9	131	17	26
XBAS	653	2.0	11.0	670	1.0	7.5	670	4	14
XQT	6940	1.1	11.6	7929	1.4	13.3	7929	5	22

Table 6-4. Key Percentile Values for All Parameters for All 300 BPS Conversations, Grouped by Application

SUBSET	Rs			Da			Ta		
	#Obs	50%	90%	#Obs	50%	90%	#Obs	50%	90%
ADD	497	5.1	30.0	488	0	0.1	497	0.2	0.4
ASG	396	2.6	4.3	396	0	0	396	0.2	1.0
ASM	7	1.9	2.5	7	0	0	7	0	0.9
BASIC	28	0.9	1.9	28	0	0	28	0.2	0.6
BRKPT	131	2.9	5.7	131	0	0	131	0.1	0.4
COB	17	3.2	30.0	17	0	0.1	17	0.2	0.2
DATA	247	30.0	30.0	239	0	0	247	0	0.2
ED	5846	30.0	30.0	5809	0	0	5846	0.2	0.6
ELT	186	30.0	30.0	186	0	0	186	0.3	0.4
FIN	227	3.2	5.8	227	0	0	227	1.3	2.7
FOR	93	2.5	30.0	93	0	0	93	0.3	1.6
FREE	215	3.1	6.1	209	0	0	215	0.2	0.5
FUR	1036	2.5	30.0	1027	0	0	1036	0.2	1.0
LIST	2	1.7	1.7	2	0	0	2	0.1	0.1
MAP	163	2.3	5.7	161	0	0	163	0.2	1.0
PMD	93	30.0	30.0	91	0	0	93	0	0.2
RALPH	20	2.5	3.9	20	0	0	20	0.2	0.9
RFOR	10	2.6	2.8	10	0	0	10	0.4	1.0
RUN	238	2.7	5.8	236	0	0.5	238	0.2	1.2
START	98	2.2	3.8	98	0	0	98	0.2	0.8
STAT	294	3.2	30.0	290	0	0	294	0.2	0.8
SYM	37	2.6	4.8	37	0	0	37	0.1	0.3
USE	131	2.3	4.2	131	0	0	131	0.2	0.4
XBAS	670	4.0	7.4	669	0	0	670	0.2	1.2
XQT	7929	3.9	30.0	7721	0	0	7929	0.2	0.5

Table 6-4. (Continued)

SUBSET	Ca			Ra			Dr		
	#Obs	50%	90%	#Obs	50%	90%	#Obs	50%	90%
ADD	455	7	7	497	30.0	30.0	206	0.0	1.0
ASG	386	7	10	396	29.0	30.0	292	0.2	0.6
ASM	7	1	2	7	2.4	30.0	7	0	0.2
BASIC	27	7	10	28	30.0	30.0	15	0	0.1
BRKPT	124	4	9	131	30.0	30.0	110	0	0.5
COB	15	7	7	17	30.0	30.0	11	0	0.4
DATA	70	3	10	247	0	30.0	183	0	0.1
ED	5593	7	10	5846	30.0	30.0	3603	0.2	0.8
ELT	185	9	9	186	30.0	30.0	60	0	0.2
FIN	226	23	36	227	21.6	30.0	129	0	3.2
FOR	91	7	10	93	11.1	30.0	63	0.2	1.9
FREE	188	7	10	215	29.0	30.0	150	0.2	0.5
FUR	964	3	11	1036	29.3	30.0	790	0	1.6
LIST	2	2	2	2	5.9	5.9	2	0	0
MAP	158	5	7	163	30.0	30.0	135	0.2	2.2
PMD	18	14	14	93	0	30.0	14	0	0
RALPH	19	2	3	20	6.9	30.0	19	0.2	0.2
RFOR	10	2	2	10	4.4	6.5	10	0.5	0.5
RUN	217	7	10	238	30.0	30.0	186	0.2	1.0
START	94	7	10	98	30.0	30.0	84	0	1.6
STAT	251	7	10	294	27.6	30.0	183	0.1	0.8
SYM	36	2	7	37	30.0	30.0	33	0	0.5
USE	126	7	10	131	30.0	30.0	74	0.1	0.5
XBAS	652	7	9	670	30.0	30.0	458	0.2	0.8
XQT	6788	7	10	7929	30.0	30.0	4472	0	1.3

Table 6-4. (Continued)

SUBSET	T^r			C^r			R^r		
	#Obs	50%	90%	#Obs	50%	90%	#Obs	50%	90%
ADD	497	0.7	10.6	451	24	299	497	30.0	30.0
ASG	396	0.5	1.5	350	16	44	396	29.9	30.0
ASM	7	0	1.3	4	30	30	7	0	23.5
BASIC	28	0	2.0	28	1	59	28	30.0	30.0
BRKPT	131	0	1.1	53	2	33	131	0	30.0
COB	17	0	3.2	15	1	95	17	30.0	30.0
DATA	247	0	1.1	219	1	35	247	30.0	30.0
ED	5846	1.2	3.7	5572	33	111	5846	30.0	30.0
ELT	186	0.1	1.5	186	2	44	186	30.0	30.0
FOR	227	1.3	2.9	215	40	67	227	29.0	30.0
FIN	93	13.4	26.3	75	403	787	93	30.0	30.0
FREE	215	0.5	1.3	193	16	40	215	30.0	30.0
FUR	1036	1.2	16.1	827	59	470	1036	28.9	30.0
LIST	2	0	0	1	-	-	2	0	0
MAP	163	1.3	4.5	131	36	125	163	28.9	30.0
PMD	93	0	5.7	86	1	196	93	30.0	30.0
RALPH	20	1.7	3.2	13	50	66	20	20.7	26.2
RFOR	10	2.0	2.6	10	44	45	10	21.6	21.6
RUN	238	1.3	2.4	213	40	71	238	30.0	30.0
START	98	0	1.2	64	1	35	98	30.0	30.0
STAT	294	6.9	22.2	287	206	622	294	30.0	30.0
SYM	37	0	1.1	17	19	42	37	0	30.0
USE	131	0.5	0.7	118	16	23	131	30.0	30.0
XBAS	670	0.8	8.7	629	26	250	670	30.0	30.0
XQT	7929	0.8	4.7	6742	30	142	7929	30.0	30.0

Table 6-4. (Continued)

SUBSET	Dsr			Dss		
	#Obs	50%	90%	#Obs	50%	90%
ADD	497	0.3	0.9	497	1.4	12.1
ASG	396	0.5	1.4	396	0.0	11.5
ASM	7	0	1.1	7	0	0
BASIC	28	0.3	0.7	28	29.8	52.3
BRKPT	131	0	1.2	131	0	9.6
COB	17	0.3	0.7	17	2.4	18.2
DATA	247	0	0.3	247	0.1	17.3
ED	5846	0.4	1.3	5846	7.4	31.9
ELT	186	0.3	0.5	186	2.4	9.2
FIN	227	1.5	5.0	227	0	0
FOR	93	1.0	2.2	93	0	13.3
FREE	215	0.3	0.9	215	0	6.4
FUR	1036	0.4	2.5	1036	1.7	37.8
LIST	2	0	0	2	0	0
MAP	163	0.5	3.8	163	4.4	15.7
PMD	93	0	0.2	93	0	18.7
RALPH	20	0.4	1.1	20	0	7.6
RFOR	10	1.0	1.5	10	0	0
RUN	238	0.7	3.0	238	0	18.5
START	98	0.4	2.1	98	0	6.7
STAT	294	0.5	1.2	294	0	50.3
SYM	37	0	0.9	37	0	5.1
USE	131	0.4	0.7	131	0	8.8
XBAS	670	0.4	2.1	670	5.6	26.0
XQT	7929	0.3	1.4	7929	4.6	28.4

Table 6-4. (Concluded)

6.4.1 Analysis Of Particular Parameters -

One simple way to analyze the data in Table 6-4 is to sort the subsets according to the values of one or more parameters. Sorting provides a quick way for the analyst to determine the range of differences between subsets. Furthermore, seeing the subsets which exhibit similar parameter values and widely different parameter values in this organized way may lead to hypotheses explaining the similarities and differences which can be tested in other ways.

One interesting pair of parameters to analyze in this way are stimulus transmission rate and stimulus character count. In Table 6-5 below, the subsets are sorted in increasing order according to their values for stimulus transmission rate (first at the 50% level and then at the 90% level where 50% values were identical). Parameter values for stimulus character count are also shown at the 50% and 90% level.

It is evident even from a cursory examination of Table 6-5 that higher stimulus transmission rates are generally associated with shorter stimuli. This is in keeping with previously reported (section 2.2.1) human factors findings that "burst" rates for short character sequences can greatly exceed average rates for longer sequences.

SUBSET		Rs		Cs	
	# Obs.	50%	90%	50%	90%
BASIC	28	0.9	1.9	20	25
LIST	2	1.7	1.7	16	16
ASM	7	1.9	2.5	14	19
START	98	2.2	3.8	19	28
USE	131	2.3	4.2	17	26
MAP	163	2.3	5.7	12	21
RALPH	20	2.5	3.9	19	34
FOR	93	2.5	30.0	16	27
FUR	1036	2.5	30.0	14	27
RFOR	10	2.6	2.8	18	18
ASG	396	2.6	4.3	15	22
SYM	37	2.6	4.8	10	21
RUN	238	2.7	5.8	31	41
BRKPT	131	2.9	5.7	14	20
FREE	215	3.1	6.1	11	17
FIN	227	3.2	5.8	5	5
COB	17	3.2	30.0	14	16
STAT	294	3.2	30.0	6	14
XQT	7929	3.9	30.0	5	22
XBAS	670	4.0	7.4	4	14
ADD	497	5.1	30.0	2	18
DATA	247	30.0	30.0	1	16
ED	5846	30.0	30.0	6	21
ELT	186	30.0	30.0	20	32
PMD	93	30.0	30.0	1	13

Table 6-5. Stimulus Transmission Rate and Stimulus Character Count by Application, Sorted by Increasing Values of Stimulus Transmission Rate

6.4.2 Analysis Of Particular Subsets -

An alternative type of analysis that can be performed with the data in Table 6-4 is to compare the characteristics of a few particular subsets. This comparison may involve all the parameters of the the model, or just a few, depending on the reason for the comparison.

For example, the author was called on some time ago to conduct a comparison between the Univac standard BASIC interpreter and XBASIC, an enhanced interpreter offered by a private software house. Of necessity, the comparison focused on qualitative features such as differences in the command vocabulary. Without considering the further implications of the results (or even their statistical validity) at this time, we simply note that the empirical data collected during this study and analyzed according to application reveal that the user think time, the stimulus transmission rate, and the stimulus character count were all considerably less for XBASIC than for BASIC, while the response character count was considerbly greater. This might suggest greater ease of use and utility for XBASIC, though obviously a more detailed investigation would be required to confirm or deny such interpretations. However, observed users who were free to choose either interpreter showed a clear preference for XBASIC as indicated by the much larger incidence of observed message groups for it.

7.0 SUMMARY AND CONCLUSIONS

In this chapter we review the results of the study as described in the previous six chapters. No new information is presented, but some conclusions based on previously presented information are offered that help to tie together some of the seemingly disparate findings. Areas of applicability for the results (particularly the methodology) are discussed, and the limitations of the current study are candidly identified. The chapter and the dissertation concludes with some suggestions for future work in this area.

7.1 Review Of Research And Findings

This study has dealt both with the development of a methodology for the quantitative evaluation of interactive computing and with the application of that methodology to a specific interactive system. In the following sections we review the methodology which is the most significant contribution of this study. The empirical results from the data collected on the particular system under test are also summarized. These results are interesting in their own right, especially when compared with previously published results.

7.1.1 Methodology -

In this dissertation we have developed and applied a new methodology for the measurement and evaluation of interactive computing. This methodology may be applied to either the users of an interactive computing system (in which case the users are considered as a system component in the traditional human factors approach) or to the interactive computing system, including the service computer and any communications network through which the service is delivered. The methodology may be used for a variety of different applications, some of which will be suggested in Section 7.2 below. Here we review the methodology itself and identify the methodological contributions of the dissertation.

The methodology is based on the application of a data collection technology developed at the National Bureau of Standards. As explained in Chapter 3, a passive recording device can be attached to the data path between the user and the computing system so that it can identify and time tag all characters flowing in either direction. In this way, all the fundamental data about interactive conversations are collected for subsequent analysis.

The basic processing software developed for the collected data applied a new model of user-computer interaction that distinguished between the "acknowledgement" and "response" portions of the computer system's transmissions. What was available when the dissertation work began was the ability to collect data, apply the model, and have certain summary statistics computed for the fourteen model parameters for various sets of message groups aggregated both within and across conversations.

As explained in section 3.1.3, the author had an option added to the DAP to write to a file the values of all the SAR model parameters for each message group in a set. Routines were then written by the author to compute cumulative frequency percentiles for each parameter, to form the distributions of the medians of the parameters for the message groups in each conversation and to compute the cumulative frequency percentiles for these distributions, to trim the distributions above or below specified data limits, to compute auto-correlation coefficients and form their distributions, and to pass the distributions of two sets of any single parameter to library programs for the Mann-Whitney and Kolmogorov-Smirnov statistical tests. These additional routines provided the analysis capability used in this study, as described in Chapter 4.

The data collection procedures and the extent of the data actually collected were described in Chapter 3. Briefly, data were collected on randomly selected days and at randomly selected times over a period of three years. Recorded conversations were culled to identify those that were representative of "normal" operation on the interactive system. These conversations were then aggregated by terminal speed and the results analyzed. Two sets of results were presented in Chapter 5 for each parameter of the SAR model:

1. Graphical and tabular presentation of the distribution of the model parameter for all observations; and

2. Tabular presentation of the distribution of conversation medians for the model parameter.

Both distributions ("all" and "medians" or "batched") for the 110 bps and 300 bps terminal classes were compared using the Mann-Whitney U-test and Kolmogorov-Smirnov test, and the results were tabulated.

The conversations recorded at 300 bps were also further aggregated by application, and a brief analysis performed according to this grouping. A complete table of results for this grouping was provided, but no statistical tests were performed.

7.1.2 Results Of Analysis By Terminal Speed -

Chapter 5 presented the results of grouping the data according to the line speed used by the interactive terminal. Data were collected for three different such speeds: 110 bps, 150 bps and 300 bps. Normalized cumulative frequency

100

distributions were formed for all the observations for each parameter, and for the median values within each conversation for each parameter.

Looking first at the parameters associated with the users' inputs to the interactive system, an increase in the input or stimulus character count was evident when all observations were considered, but not when the distributions of conversation medians were examined. This reflected the effects of of high serial correlation for this parameter, and indicated that a subset of users tended to have long system inputs throughout their conversations.

The user rate of data entry (burst rate, measured from when the input began) increased with the increase in terminal speed. This trend became clearest when a correction was made for paper tape input and single character inputs, which were transmitted at the maximum data rate and thus distorted the distributions. Comparison with previously reported data (described in section 2.2.1), particularly the Bell Laboratories' data (Jackson and Stubbs, 1969) showed a higher user data rate for NBS users at 110 bps, a slightly lower data rate at 150 bps (though the number of observations was small), and a higher data rate at 300 bps than any reported previously (though the previous study did not include the 300 bps case). The generally higher user data rates found in the present study may reflect a more rapid pace of interaction due to the extremely good response of the interactive system. This would be in keeping with the previous data that showed a lower user data rate on a more heavily loaded system (Table 2-1).

The response-stimulus delay time, or the user "think" time, was found to decrease with increased terminal speed (after the effects of serial correlation in the observations was removed by batching), indicating an increase in pace on the part of the users of higher speed terminals.

Looking at the computer output, the length of the responses were found to increase significantly with terminal speed. Thus, it is evident that terminal speed is a bottleneck for computer output, and that users will adapt their behavior in terms of the types of responses they request according to the speed of the terminal used. (We must conclude that it is the users, not the computer system, who are adapting to the speed of the terminal, since the software providing the response is the same regardless of the terminal speed.

If we can make the assumption that users are more satisfied with longer responses (since they request such responses when they are able), we can make an interesting comparison with other data reported by L. H. Miller (described in Section 2.1.5). Miller found no significant performance or attitude change associated with an increase in terminal speed from 120 cps to 240 cps. We found a significant increase in input and output volume (with a presumed increase in user satisfaction, though user attitudes were not measured directly) with an increase in terminal speed from 10 cps to 30 cps.* This leads us to suggest the relationship between terminal speed and user satisfaction shown in Figure 7-1. We suggest that an elbow in the curve must exist somewhere between 30 cps and 120 cps. Further experiments with a wider range of terminal speeds would be required to locate it precisely.

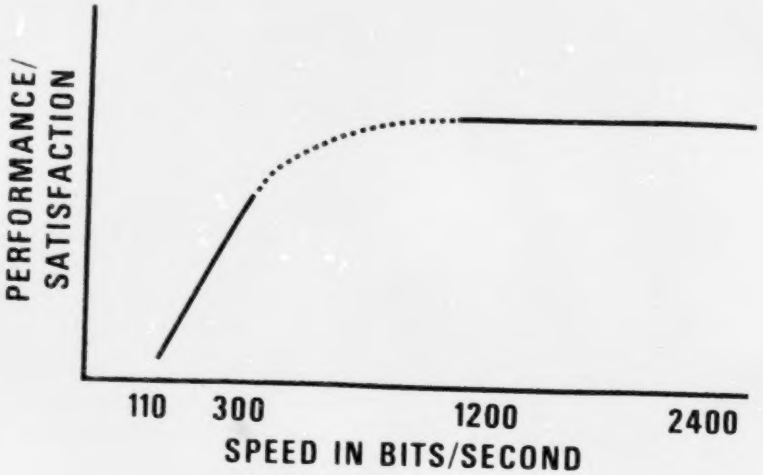


Figure 7-1. User Satisfaction as a Function of Terminal Speed

Response transmission rate is one parameter for which the mean values are more enlightening than the percentiles. Response transmission rate clearly increases with terminal speed; the median values seem to indicate that responses were always transmitted at the maximum rate. However, the mean values (from the line utilization statistics) were considerably less, reflecting the effects of a moderate number of relatively slowly transmitted responses. Much of these delays can probably be attributed to output queue processing in the computer system after it has begun the transmission of long responses. The idle time occuring during the response portion of the conversation represent opportunities to regain the use of otherwise wasted transmission capacity through some line sharing technology such as statistical multiplexing or concentration. This study highlights the opportunities for sharing that exist even on relatively slow transmission lines.

---

* Differences between the distributions of all observations of stimulus character count and response character count for the 110 bps class and the 300 bps classes were significant at the 1% level; however, the differences between the distributions of conversation medians for the same parameters and classes were not significant at that level. Stimulus character count seemed to exhibit considerable serial correlation (Figure 4-1-c) while response character count did not (Figure 4-1-k).

Response delay times were found to be slightly longer for higher speed terminals; however, the level of the system response to the users was so good in all cases, and the differences so imperceptible as to make this finding relatively unimportant.

The acknowledgement portion of the message group was also found to be insignificant for the interactive system under investigation. This can be attributed to the relatively light interactive load on the system during the study period. During this period, the impact of interactive users on the system was being evaluated and access was limited to a small number of simultaneous users. In retrospect, it seems clear that the simpler S-R model could indeed have been used for this study without impairing the results. This in no way invalidates the SAR model, but rather demonstrates that the added detail (and complexity) is not necessary in all cases.

7.1.3 Results Of Analysis By Application -

Chapter 6 presented the results of grouping the data by user application prior to analysis (only data collected for 300 bps terminals was used, so that terminal speed was not a factor in this analysis). The 50% and 90% values of all parameters are presented for all subsets; however, only some of the data presented was discussed.

Examples were given of two types of analyses: analysis of particular parameters and analysis of particular subsets. In the analysis of particular parameters, the stimulus transmission rate and stimulus character count were examined for all application subsets. The subsets were sorted by stimulus transmission rate, and then both parameters were displayed in tabular form. This permitted easy recognition of which applications had characteristically higher or lower stimulus transmission rates. It was also easy to notice that high stimulus transmission rates were generally associated with shorter stimulus character counts, and vice versa.

In the analysis of particular subsets, it was demonstrated how two similar language processors (BASIC and XBASIC were the two chosen for the example) could be compared quantitatively. Since the intent was only to illustrate additional applications for the methodology, statistical testing was not performed for the two classes of data.

7.2 Applicability Of Results

The general approach that has been developed in this dissertation should be useful for a variety of applications. In particular, the ability to test for significance of differences between sets of non-normally distributed data is essential if objective comparisons are to be made between performance variables in interactive computing. The following sections present several examples of possible areas of application for the methodology.

7.2.1 Procurement -

The analysis techniques which have been developed may be used in the design, selection, procurement and improvement of computer systems and remotely provided computer services. In the design of a computer system or network, the ability to measure external performance is useful in specifying service design goals and in determining how well these goals are met (Kamerman, 1969). Objective measurement will allow the direct comparison of alternatives by implementers and users alike. Thus, service providers can implement systems with specific design goals for service, and users can specify, perhaps contractually, the level of service they expect (Grubb and Cotton, 1975).

Specification of the anticipated workload is an essential part of any request for proposal (Ferrari, 1972). Knowledge of current demand provides a starting point for forecasting future demand by providing a profile of user characteristics. Knowledge of current responsiveness, together with current user evaluations, provides the basis for strengthening or relaxing such requirements in the future. Both buyer and seller should benefit from having a straightforward way to determine if contract terms are being met. Finally, these measurement and analysis techniques can be applied to installed systems to determine the level of service being provided, identify where improvements are needed, and to evaluate the effects of changes made to the system.

Application of the methodology to proposed systems could be part of any benchmarking evaluation to determine if responsiveness is as claimed in the bid.* After selection and procurement, the methodology could continually be used to ensure compliance with contractual provisions. This would require that performance requirements be specified in the terms of the contract -- a practice that is not too common today. The statistical tools developed as a part of this dissertation could be used to demonstrate non-compliance with contractural service specifications.

---

* In fact, similar techniques (based on the same measurement device used in this study) have already been applied in the procurement process (Abrams and Hayden, 1978), but without the full statistical rigor of the methodology developed here.

7.2.2 User Parameters For System Design -

Considerable data characterizing terminal users has been collected. Knowledge of the distributions for such parameters as typing rate and length of message are essential to the design of modern computer communications systems. Most current communications networks and timesharing systems are designed for shared use under the principle that users are not active all the time. The number of interactive users who can "simultaneously" be served in this fashion (without extreme degradation of service) is a statistical function of the behavior patterns of the users.

Descriptive data has been obtained for a sample population of scientific timesharing users employing modern terminals. This data has been compared with widely circulated data collected by Jackson, Stubbs and Fuchs (Jackson and Stubbs, 1969; Fuchs and Jackson, 1970) and differences have been noted. However, neither of these sets of data should be taken as representative of all interactive applications. Designers of systems for particular applications could to employ a similar methodology to collect data descriptive of users engaged in tasks more typical of their application.

Other types of user-oriented factors could also be investigated with these techniques. For example, the error conditions that led to many of the recorded conversations being discarded could be analyzed in detail. With greater control and/or knowledge of the users being measured, a variety of human factors experiments can be conducted, including investigation of the effects of user experience (Walther and O'Neil, 1974), user learning rate (Jutila and Baram, 1971), or different terminal types (Walther and O'Neal, 1974).

7.2.3 Tuning Interactive Systems -

System tuning refers to the practice of making small changes in system hardware or software in the effort to optimize system performance (such as by eliminating bottlenecks). For tuning interactive systems, installation managers need a way to assess the impact of system modifications on service to users. These techniques can be applied to evaluating in a quantitative way the effects of system changes on system performance. By taking measurements before and after a system modification which is expected to improve performance, the level of improvement that is realized can be measured. Thus, an objective measure of service improvement can be provided to installation management for their use in determining whether the resources expended were justified by the results realized.

The evaluation techniques can be applied for either hardware or software changes, and even to determine if software improvement efforts seem warranted. By measuring the different subsystems (e.g., the various language compilers) within a single system, differences in levels of service can be detected. The service characteristics of the different subsystems can be expected to have an effect on their value to users, and may help to explain user preferences. This type of information should also be of value to installation management in allocating available resources for user training and for software maintenance and improvement.

Similarly, knowledge of the relative utilization of various facilities by users can provide guidance as to where optimization efforts should be directed.

7.3 Limitations

In view of the differences that have been noted between the results of this study and the results of studies with terminals at far greater transmission speeds (L. H. Miller, 1976, 1977), it is clear that the relatively narrow range of terminal speeds studies is one limitation of the study. It would be desirable to perform a study that applied a consistent methodology to a wide range of terminal speeds, so that conclusive results could be obtained.

A second limitation is that the measurement approach described here offers no direct explanations as to the level of service provided by the computer system being tested. All the measurements are strictly external to the system under investigation. While these measurements could be correlated with measurements taken internally (e.g., from software monitors) in order to gain some additional insight as to the functioning of the host, no such correlations are proposed here.

Also, no attempt is made to group users according to any schema such as experience with the system or time of day of usage. Rather, the experimental design called for large samples of typical usage. In this way, the peculiarities of any individual user should not have had any significant effect on the results.

Reliance on response time as the primary measure of service could be viewed as a limitation of the study. (For example, Abrams and Treu (1977) identified more than fifty different measures relating to user-computer interactive behavior). However, the data collected in the study can be used to investigate a number of different service measures. Furthermore, some selection had to be made as to which measures to consider in order to provide bounds to the study.

A final comment is appropriate regarding the measurement system employed. This is currently a unique tool only available within the Federal Government, where these techniques are currently being recommended (National Bureau of Standards, 1978).

However, this measurement approach could be replicated with current technology at nominal cost. The programs that implement the SAR model are written in Fortran and should be portable. Alternatively, they could be rewritten for a different system from the algorithms developed. Thus, application of the methodology elsewhere is feasible.

7.4 Suggestions For Future Work

The type of data collected and analyzed in this study is rich with possibilities for further research. By providing a methodology to collect and analyze large volumes of data on basic events (such as the transmission of individual characters and character strings by the users and the computer), many aspects of interactive computer usage that were previously only dealt with in qualitative terms may be analyzed quantitatively.

A number of the limitations just observed could be remedied by subsequent studies that covered a wider range of terminal speeds or controlled more variables. We have already noted the desirability of correlating internal measurement data with external data of the sort reported here. Studies which repeatedly recorded data from a single or known group of users could identify trends over time as well as differences between individuals or between groups. By randomly forming groups from a homogeneous population and then varying one or more of the experimental conditions (such as terminal type in addition to terminal speed) a wide variety of different experimental questions can be investigated.

It would also be quite interesting to try to combine this type of quantitative data with qualitative, such as information on user-perceived service quality (Dzida, Herda and Itzfeldt, 1978). This would serve to validate the use of these quantitative measures as determinants of interactive performance. The wide range of applications described above should also serve to suggest future work based on the methodology developed here.

BIBLIOGRAPHY

1. Abrams, M. D., G. E. Lindamood, and T. N. Pyke, Jr., "Measuring and Modeling Man-Computer Interactions," *Proceedings of the 1st Annual SIGME Symposium on Measurement and Evaluation*, February 1973, pp. 136-142.

2. Abrams, M. D., and I. W. Cotton, *The Service Concept Applied to Computer Networks*, National Bureau of Standards, Technical Note 880, August 1975.

3. Abrams, M. D., I. W. Cotton, S. W. Watkins, R. Rosenthal, and D. E. Rippy, "The NBS Network Measurement System," *IEEE Transactions on Communications*, *Vol. COM-25*, *No. 10*, October 1977, pp. 1189-1198.

4. Abrams, M. D. and S. Treu, "A Method for Interactive Computer Service Measurement," *Communications of the ACM*, December 1977, pp. 936-944.

5. Abrams, M. D. and H. P. Hayden, "Application of a Network Monitor to the Selection of a Time Shared Computing System," *Computer Performance Evaluation Users Group (CPEUG) 14th Meeting*, National Bureau of Standards Special Publication 500-41, October 1978, pp. 15-25.

6. Amer, P. D. and S. A. Mamrak, "Statistical Methods in Computer Performance Evaluation: A Binomial Approach to the Comparison Problem," *Computer Science and Statistics: Eleventh Annual Symposium on the Interface*, Institute of Statistics, North Carolina State University, March 1978, pp. 314-319.

7. Bell, T. E., "Computer performance variability," *National Computer Conference*, 1974, pp. 761-766.

8. Boehm, B. W., M. J. Seven, and R. A. Watson, "Interactive Problem Solving -- An Experimental Study of Lockout Effects," *Spring Joint Computer Conference*, 1971, pp. 205-210.

9. Boies, S. J., "User Behavior on an Interactive Computer System," *IBM System Journal*, *Vol. 13*, *No. 1*, 1974, pp. 2-18.

10. Bryan, G. E., "JOSS: 20,000 hours at a console -- a statistical summary," *Fall Joint Computer Conference*, 1967, pp. 769-777. 500-18, September 1977, pp. 107-111.

11. Carbonell, J. R., J. I. Elkind and R. A. Watson, "On the Psychological Importance of Time in a Time Sharing System," *Human Factors*, *Vol. 10*, No. 2, 1968, pp. 135-142.

12. Cotton, I. W., Review of "Response Time in Man-Computer Conversational Transactions" by R. B. Miller, Review No. 17,744, Computing Reviews, October 1969, p. 498.

13. Cotton, I. W., "Cost-Benefit Analysis of Interactive Systems," Proceedings of 2nd Jerusalem Conference on Information Technology, August 1974, pp. 729-746.

14. Davies, D. R. and G. S. Tune, Human Vigilance Performance, New York: American Elsevier Publishing Company, 1969.

15. Devoe, D. B., "Alternatives to Handprinting in the Manual Entry of Data," IEEE Transactions on Human Factors in Electronics, January 1967, pp. 21-32.

16. Dudick, A. L., E. Fuchs and P. E. Jackson, "Data Traffic Measurements for Inquiry-Response Computer Communications Systems," Proceedings of the IFIP, Ljubljana, Yugoslavia, August 1971, pp. 634-641.

17. Dzida, W., S. Herda and W. D. Itzfeldt, "User- Perceived Quality of Interactive Systems, IEEE Transactions on Software Engineering, July 1978, pp. 270-276.

18. Ferrari, D., "Workload Characterization and Selection in Computer Performance Measurement," Computer, July/August 1972, pp. 18-24.

19. Fuchs, E. and P. E. Jackson, "Estimates of Distributions of Random Variables for Certain Computer Communications Traffic Models," Communications of the ACM, Vol. 13, No. 12, December 1970, pp. 752-757.

20. Gibbons, J. D., I. Olkin and M. Sobel, Selecting and Ordering Populations: A New Statistical Methodology, New York: John Wiley & Sons, 1977.

21. Grubb, D. S. and I. W. Cotton, Requirements of Information Processing Systems for Supporting Telecommunications Services, NBS Technical Note 882, September 1975.

22. IBM System/360 Scientific Subroutine Package, Version III, Programmers Manual, IBM Data Processing Division, White Plains, New York, Report GH 20-0205-4, 1968.

23. Jackson, P. E. and C. D. Stubbs, "A Study of Multiaccess Computer Communications," Spring Joint Computer Conference, 1969, pp. 491-504.

24. Jutila, S. T. and G. Baram, "A User-Oriented Evaluation of a Time-Shared Computer System," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 1, No. 4, October 1971, pp. 344-349.

25. Kamerman, A., Establishing and Measuring Conversational Performance Objectives for Time Sharing Systems, Technical Report TR53.015, IBM Systems Development Division, Yorktown Heights, N.Y., June 1969.

26. Klemmer, E. T. and G. R. Lockheed, "An Analysis of Productivity and Errors on Keypunches and Bankproof Machines," IBM Research Report, RC-354, IBM Research Center, Yorktown Heights, N.Y., November 1960.

27. Lancaster, F. W. and E. G. Fayen, Information Retrieval On-Line, Los Angeles: Melville Publishing Co., 1973.

28. Mackworth, J. F., Vigilance and Attention: A Signal Detection Approach, Middlesex, England: Penguin Books, 1970.

29. Mamrak, S. A. and P. A. DeRuyter, "Statistical Methods for Comparing Computer Services," Computer, November 1977, pp. 32-39.

30. Mamrak, S. A. and P. D. Amer, "A Methodology for the Selection of Interactive Computer Services," National Bureau of Standards Technical Note, 1978 (in press).

31. Mann, H. B. and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," Annals of Mathematical Statistics, 18, 1947, pp. 50-60.

32. Miller, L. A. and J. C. Thomas, Jr., "Behavioral Issues in the Use of Interactive Systems," Research Report RC-6326, IBM Research Laboratory, Yorktown Heights, N.Y., December, 1976.

33. Miller, L. H., "An Investigation of the Effects of Output Variability and Output Bandwidth on User Performance in an Interactive Computer System," University of Southern California, Information Sciences Institute, Technical Report 76-50, December 1976.

34. Miller, L. H., "A Study in Man-Machine Interaction," National Computer Conference, 1977, pp. 409-421.

35. Miller, R. B., "Response Time in Man-Computer Conversational Transactions," Fall Joint Computer Conference, 1968, pp. 267-277.

36. Mostofsky, D. I. (ed.), Attention: Contemporary Theory and Analysis, New York: Appleton-Century-Crofts, 1970.

37. National Bureau of Standards, Guidelines for the Measurement of Interactive Computer Service Throughput and Response Time, Federal Information Processing Standards Publication 57, August 1978.

38. Rosenthal, R., D. E. Rippy and H. Wood, *The Network Measurement Machine -- A Data Collection Device for Measuring the Performance and Utilization of Computer Networks*, NBS Technical Note 912, April 1976.

39. Seibel, R., "Data Entry Devices and Procedures," In *Human Engineering Guide to Equipment Design*, Harold P. Van Cott and Robert G. Kinkade, editors. Sponsored by the Joint Army-Navy-Air Force Steering Committee, Washington, D.C.: U.S. Government Printing Office, 1972.

40. Siegel, S., *Non-parametric Statistics*, New York: Mc-Graw Hill, 1956.

41. Smirnov, N. V., "Table for estimating the goodness of fit of empirical distributions," *Annals of Mathematical Statistics*, *19*, 1948, pp. 279-281.

42. Sobel, M., "Nonparametric Procedures for Selecting the t Populations With the Largest -Quantiles," *Annals of Mathematical Statistics*, *Vol. 38*, 1967, pp. 1804-1816.

43. Van Cott, H. P. and M. J. Warrick., "Man as a System Component," In *Human Engineering Guide to Equipment Design*, Harold P. Van Cott and Robert G. Kinkade, editors. Sponsored by the Joint Army-Navy-Air Force Steering Committee, Washington, D.C.: U.S. Government Printing Office, 1972.

44. Walther, G. H. and H. F. O'Neil, Jr., "On-line User-Computer Interface -- The Effects of Interface Flexibility, Terminal Type, and Experience on Performance," *National Computer Conference*, 1974, pp. 379-384.

45. Watkins, S. W. and M. D. Abrams, *Interpretation of Data in the Network Measurement System*, NBS Technical Note 897, February 1976.

*U.S. GOVERNMENT PRINTING OFFICE : 1979 0-281-067/146

NBS-114A (REV. 9-78)

U.S. DEPT. OF COMM.
BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION OR REPORT NO.
SP 500-48

2. Gov't. Accession No.

3. Recipient's Accession No.

4. TITLE AND SUBTITLE
Computer Science & Technology:
Measurement of Interactive Computing: Methodology and Application

5. Publication Date
June 1979

6. Performing Organization Code

7. AUTHOR(S)
Ira W. Cotton

8. Performing Organ. Report No.

9. PERFORMING ORGANIZATION NAME AND ADDRESS

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, DC 20234

10. Project/Task/Work Unit No.

11. Contract/Grant No.

12. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS *(Street, City, State, ZIP)*

13. Type of Report & Period Covered
NA

14. Sponsoring Agency Code

15. SUPPLEMENTARY NOTES
Library of Congress Catalog Card Number: 79-600079

☐ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

16. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)*

This dissertation addresses the measurement of interactive computing, including both the computer system providing service and the users demanding and receiving it. The focus is on the performance of the user and the system in individual interaction sessions (rather than on the performance of the system under varying conditions of load). A new measurement tool developed at the National Bureau of Standards is employed to record a large number of individual interactive sessions over a period of three years. The basic data of interest are the number and rate of characters sent by user and system, and latencies or delays prior to and during transmission by either party. These data are fit to a model of user-computer interaction which distinguishes between stimuli from the user, acknowledgements from the system (which only indicate that a service request has been received) and responses from the system (which contain meaningful information).

Analysis of the data consists of grouping according to two independent criteria: 1) maximum operating line speed of the terminal (either 10, 15 or 30 characters per second); and 2) type of application (for each individual service request). The data are grouped according to these criteria and cumulative frequency distributions are computed for each of 14 parameters of the model. Non-parametric tests are used to determine the significance of differences in the distributions of different sets of data.

The methodology itself is the major contribution of the study, providing, as it does, a quantitative way to investigate a variety of phenomena associated with interactive computing. The most interesting specific finding from the data collected is the increase in output data length as the terminal speed increases.

17. KEY WORDS *(six to twelve entries; alphabetical order; capitalize only the first letter of the first key word unless a proper name; separated by semicolons)*
Computer performance evaluation; human factors; interactive computing; man-machine interaction; performance measurement; timesharing.

18. AVAILABILITY ☐ Unlimited

☐ For Official Distribution. Do Not Release to NTIS

☐ Order From Sup. of Doc., U.S. Government Printing Office, Washington, DC 20402, SD Stock No. SN003-003-02081-8

☐ Order From National Technical Information Service (NTIS), Springfield, VA. 22161

19. SECURITY CLASS (THIS REPORT)
UNCLASSIFIED

20. SECURITY CLASS (THIS PAGE)
UNCLASSIFIED

21. NO. OF PRINTED PAGES
111

22. Price
$4.00

USCOMM-DC